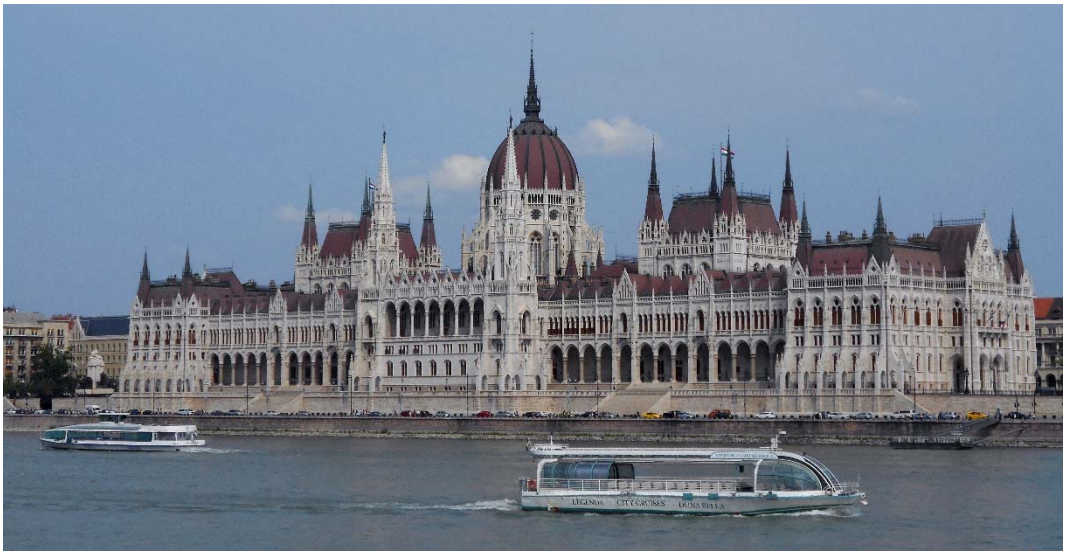
Proceedings of

DiSS 2019

The 9th Workshop on Disfluency in Spontaneous Speech

ELTE Eötvös Loránd University
Budapest, Hungary
12–13 September, 2019



ISBN: 978-963-489-063-8

Edited by
Ralph L. Rose & Robert Eklund

<This page intentionally left blank>

Proceedings of

DiSS 2019

The 9th Workshop on Disfluency in Spontaneous Speech

ELTE Eötvös Loránd University
Budapest, Hungary
12–13 September, 2019

ISBN: 978-963-489-063-8

Edited by
Ralph L. Rose & Robert Eklund

Conference website: http://diss2019.elte.hu/

Cover design by Ralph L. Rose & Robert Eklund
Graphics and photographs by Judit Bóna (except ISCA and ELTE logotypes)
Proceedings of DiSS 2019, The 9th Workshop on Disfluency in Spontaneous Speech
Workshop held at the Eötvös Loránd University (ELTE), Budapest, Hungary, 12–13 September, 2019
Editors: Ralph L. Rose & Robert Eklund
ELTE Faculty of Humanities
Eötvös Loránd University (ELTE)
H-1088 Budapest, Múzeum krt. 4., 6–8, Hungary
ISBN: 978-963-489-063-8
DOI: https://doi.org/10.21862/diss-09

Table of contents

Applied linguistics (interpreting and clinical linguistics)

Special day on (dis)fluency in children's speech

Committees

Organizing Committee

Judit Bóna, Chair
ELTE, Hungary

Márton Bartók
ELTE, Hungary

Andrea Deme
ELTE, Hungary

Robert Eklund
Linköping University, Sweden

Alexandra Markó
ELTE, Hungary

Vered Silber-Varod
The Open University of Israel

Valéria Krepsz
HAS, Hungary

Viola Váradi
ELTE, Hungary

Organizers of the Special Day (Child Language Research Group, ELTE)

Judit Bóna
ELTE, Hungary

Mária Gósy
RIL HAS and ELTE, Hungary

Viktória Horváth
RIL HAS, Hungary

Ágnes Jordanidisz
NILD Hungary

Tímea Vakula
ELTE, Hungary

Viola Váradi
ELTE, Hungary

Proceedings

Ralph L. Rose
Waseda University, Japan

Robert Eklund
Linköping University, Sweden

International Scientific Committee

Mária Gósy, Chair
RIL HAS and ELTE, Hungary

Liesbeth Degand
Université catholique de Louvain, Belgium

Jens Allwood
Gothenburg University, Sweden

Mária Bakti
University of Szeged, Hungary

Martin Corley
University of Edinburgh, UK

Jens Edlund
KTH Speech, Music and Hearing/Centre for Speech Technology, Stockholm, Sweden

Robert Eklund
Linköping University, Sweden

Keelan Evanini
Educational Testing Service, USA

Dorottya Gyarmathy
Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary

Robert Hartsuiker
Ghent University, Belgium

Viktória Horváth
Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary

Robin Lickley
Queen Margaret University, Scotland

Marianne Pouplier
Ludwig-Maximilans Universität, Germany

Ralph L. Rose
Waseda University, Japan

Elizabeth Shriberg
SRI International, USA

Vered Silber-Varod
The Open University of Israel,

Michiko Watanabe
National Institute for Japanese Language and Linguistics, Japan

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), Edinburgh (2015) and Stockholm (2017), the organizers are proud to present DiSS 2019, held at the ELTE Eötvös Loránd University, Budapest, Hungary, in September 2019.

This year, as before, we have many different approaches to disfluencies. Papers are presented in the field of both fundamental and applied research. A special feature of the 2019 event is the co-located special day focusing on (dis)fluencies in children's speech.

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 support. Special thanks to the ELTE Eötvös Loránd University, the Hungarian Linguistics Society, and the Hungarian National Research, Development and Innovation Office of Hungary (project No. K-120234).

Budapest, September 2019

Judit Bóna

Processing disfluencies in distinct speaking styles: Idiosyncrasies and transversality

Helena Moniz
School of Arts and Humanities, University of Lisbon, Lisbon, Portugal
Spoken Language Systems Laboratory, INESC-ID, Lisbon, Portugal

This talk will tackle the idiosyncratic properties of disfluencies in distinct speaking styles, mostly university lectures (Trancoso et al., 2008) and map-task dialogues (Trancoso et al., 1998), but also featuring verbal fluency tests, and (more recently) second language learning presentations in ecological settings. It will also discuss the transversal acoustic-prosodic properties pertained across speaking styles. The main research questions are twofold: i) are there domain effects in the production of disfluencies when speakers adjust to distinct communicative contexts, as in university lectures and dialogues?; ii) if domain effects do exist, are there still acoustic-prosodic properties that can be shared across domains?

As for speaking style effects in the production of disfluencies (Moniz et al., 2014), results show that there are statistical significant differences in the acoustic-prosodic parameters when speakers adjust to distinct communicative contexts. Although there is a statistically significant cross-style strategy of prosodic contrast marking (pitch and energy increases) between the region to repair and the repair of fluency, this strategy is displayed differently depending on the specific speech task, with a stronger prosodic contrast marking of disfluency-fluency repair on university lectures. In this respect, disfluencies can also be considered as a feature of a charismatic university teacher. The speaker perceived as more fluent monitors the range of energy and f0 slopes from the *reparandum* to the repair of fluency, showing also more systematic temporal measures, mirroring the behaviour of spontaneous dialogues and further enriching the class with the dynamics of a spontaneous dialogue.

Building upon the linguistic analysis, automatic speech processing experiments will also be described aiming at shedding light on the impact of the idiosyncrasies/transversality of the disfluencies. One of those experiments discriminates between list effects, disfluencies, and other linguistic events in an animal naming task (Moniz et al., 2015a). Recordings from 42 Portuguese speakers were automatically recognized and AuToBI (Rosenberg, 2010) was applied in order to detect prosodic patterns, using European Portuguese and English models. Both models allowed to differentiate list effects from the other events, mostly represented by the tunes: L* H/L(−%) (English models) or L*+H H/L(−%) (Portuguese models) contrasting with the *plateau* of the most frequent disfluency in the corpus, filled pauses. However, English models proved to be more suitable because they rely on substantial more training material.

Results on cross-domain experiments and the robustness of acoustic-prosodic features will be presented (Moniz et al., 2015b). The main trend found is that models can be quite robust across corpora for this task, despite their distinct nature. The model trained on dialogues proved to be the more robust one, possibly due to the fact that dialogues contain more contrastive tempo characteristics, while sharing with university lectures most of the pitch and energy patterns on disfluent sequences. Therefore, a model created with such data generalizes better.

Recently, cross-domain analysis of disfluencies has also been tackled in a holistic view, i.e. discussing how distinct/similar disfluencies are in the discourse markers ecosystem (Cabarrão et al., 2018). The study shows that turn-initial discourse markers are usually easier to classify than disfluencies. Our in-domain experiments achieved an accuracy of about 87% in university lectures and 84% in dialogues. The results for cross-domain are about 11%–12% lower, but still the data from one domain can be used to classify the same events in the other. Ultimately, using exclusively acoustic-prosodic cues, discourse markers can be fairly discriminated from disfluencies and sentence-like units.

In order to better understand the contribution of each feature, we have also reported the impact of the features in both the dialogues and the university lectures. Pitch features are the most relevant ones for the distinction between discourse markers and disfluencies, namely pitch slopes. Furthermore, although they have idiosyncratic properties, disfluencies, particularly filled pauses, may share with discourse markers the same prosodic properties, as the plateau contours contrasting with rises in the following prosodic constituents.

Future challenges will encompass human-computer interactions both with virtual and embodied agents aiming at simulating both the

https://doi.org/10.21862/diss-09-001-moniz

idiosyncratic traits of the domains and the shared acoustic-prosodic features across such domains.

Acknowledgements

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, and under Ph.D. grant SFRH/BD/96492/2013, and Post-doc grant SFRH/PBD/95849/2013. The author is very grateful to work with her co-authors and friends, Vera Cabarrão, Fernando Batista, Rubén Solera Ureña, Jaime Ferreira, Ana Isabel Mata, Isabel Trancoso. Furthermore, the author is also very grateful for discussions and feedback provided by several researchers, with a special thank you for Julia Hirschberg, Ralph Rose, and Vered Silber-Varod.

References

Cabarrão, V., H. Moniz, F. Batista, J. Ferreira, I. Trancoso & A. I. Mata. 2018. Cross-domain Analysis of Discourse Markers in European Portuguese. *Dialogue & Discourse* 9(1): 79–106.

Moniz, H., F. Batista, A. I. Mata & I. Trancoso. 2014. Speaking Style Effects in the Production of Disfluencies. *Speech Communication* 65: 20–35. https://doi.org/10.1016/j.specom.2014.05.004

Moniz, H., A. Pompili, F. Batista, I. Trancoso, A. Abad & C. Amorim. 2015a. Automatic Recognition of Prosodic Patterns in Semantic Verbal Fluency Tests – an Animal Naming Task for Edutainment Applications. In: The Scottish Consortium for ICPhS 2015 (ed.), *Proceedings of The 18th International Congress of Phonetic Sciences* (ICPhS 2015), 10–14 August 2015, Glasgow, Scotland, paper number 0997.1-5.

Moniz, H., J. Ferreira, F. Batista & I. Trancoso. 2015b. Disfluency Detection across Domains. In: *DiSS 2015, Proceedings of the 7th Workshop on Disfluency in Spontaneous Speech*, 8-9 August 2015, University of Edinburgh, Scotland, UK.

Rosenberg, A. 2010. AuToBI – A Tool for Automatic ToBI Annotation. In: T. Kobayashi, K. Hirose & S. Nakamura (eds.), *Proceedings of the 11th Annual Conference of the International Speech Communication Assocation* (Interspeech 2010), 26–30 September 2010, Makuhari, Chiba, Japan, 146–149.

Trancoso, I., M. C. Viana, I. Duarte & G. Matos. 1998. Corpus de Diálogo CORAL. In: *Proceedings of Proceesing of the Portuguese Language* (PROPOR 1998), 3–4 November 1998, Portalegre, Brazil.

Trancoso, I., R. Martins, H. Moniz, A. I. Mata & M. C. Viana. 2008. The LECTRA Corpus – Classroom Lecture Tanscriptions in European Portuguese. In: Nicoletta Calzolari (Conference Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC 2008), 28–30 May 2008, Marrakesh, Morocco, 1416–1420.

Halt command in word retrieval

Mária Gósy
Research Institute for Linguistics, Hungarian Academy of Sciences, and
Department of Phonetics, ELTE University, Budapest, Hungary

Abstract

In this study, occurrences and temporal patterns of five types of disfluencies were analyzed that show a common feature on the surface. All of them have some kind of interruption of content words followed by some continuation. The purpose was to show whether the place of interruption of the word articulation and the durational patterns of the editing phases are characteristic of re-starts, false starts, slips of the tongue, pauses within words, and prolongations. More than 1,400 instances were processed. Both (i) the number of pronounced segments of abandoned words and the duration of the corresponding editing phases are characteristic of a specific disfluency type, and (ii) speakers select a strategy to overcome their speech planning difficulties most economically.

Introduction

A wide variety of errors—like conceptual errors, syntactic errors, lexical errors, phonemic/phonetic (Ph-) errors, prosodic errors, morphemic errors, errors relating to social context (etc.)—and other types of disfluencies—like filled pauses, prolongations, repetitions (etc.)—can occur when people speak spontaneously (Levelt, 1989; Postma, 2000). Although the types and occurrences of disfluencies are different across speakers, speech styles, and languages, practically all speakers produce them. The speech monitor is a well-developed mechanism that is sensitive to several processes and their outcomes during speech planning and execution, and makes it possible for the speaker to prevent, detect and repair real errors. Although there is no accomplished consensus concerning the nature and character of the process of internal monitoring (e.g. Levelt, 1989; Huettig & Hartsuiker, 2010), everybody accepts the existence of such a monitor having a definite task during speech production. Real erroneous words and those that are judged by the speaker to be erroneous are sometimes interrupted on the surface as a result of a 'halt command' the articulatory organs receive. The halt command is preceded by some discrepancy of intended and monitored speech calling for some context-dependent solution.

Self-initiated self-interruptions are defined as the halting of the continuous articulation of a word at a certain point. These interruptions follow the classical process: A speaker interrupts the flow of speech at an 'interruption point' to repair (or modify) something s/he has said before, and/or monitors the produced speech prior to halting (Levelt, 1989; Nooteboom & Quené, 2008). The stretch of speech to be repaired is the 'reparandum' while the correction is the 'repair'. The 'repair' sometimes does not contain any change or modification, like in the case of re-starts. The editing phase is frequently a silent period or a filled pause, or may contain cue phrases, or any combination of these. The editing phase can be completely missing when the repair immediately follows the reparandum. In this study we wanted to find answers for the interrelations between certain disfluencies and interruption points (i.e. the number of pronounced segments before halting) as well as between these disfluencies and the durations of their editing phases. Hungarian is an agglutinating language, so content words can easily consist of 10 or more segments (frequently more than 5 syllables) due to suffixation, a fact that provides an excellent possibility to analyze the various numbers of segments before halting.

In the line of Blackmer and Mitton's (1991) view we claim that many disfluencies are edited errors. Edited errors concern Ph-errors and false starts where the errors appear on the surface and should be repaired. Although the notion of editing phase is rather well defined in the literature, we will interpret it in a broad sense as any period between the interruption of word articulation and its continuation, whether it is implemented as silent pause, filled pause, some cue phrase, nothing (immediate continuation), or even some unexpected prolongation of a segment.

There are disfluency phenomena where both reparandum and repair are assumed to exist, but no surface error can be detected. Editing phases in these cases are supposed to refer to monitoring and finding a strategy for problem solution during speech planning. These disfluency phenomena are (among others) re-starts, pauses within words (PwW) and prolongations. In the case of re-starts there is no erroneous part of the word, but the speaker senses a possible error during monitoring. The monitoring and the supposed preparedness for repair require extra time resulting in an editing phase between halting and continuation. Monitoring terminates with

https://doi.org/10.21862/diss-09-002-gosy

the conclusion that no error could be found. So, the pronounced part of the word will be repeated, and word retrieval and its articulation completed. A very similar process takes place in the case of the phenomenon of PwW. The speaker interrupts word articulation and after a while there is a continuation resulting in completing the word. The difference between re-starts and PwWs is the lack of repetition of the reparandum in the PwW cases. We suggest that there should be a third variant of this process with an editing phase that manifests itself in segment prolongation. Interruptions of the speech flow within a word may refer to various problems that exist in any phase of speech planning, lexical access and pronunciation. Figure 1 illustrates our broadened conception of the editing phase.

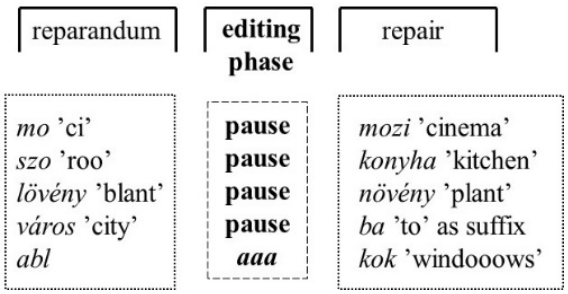


Figure 1. The structure of error, interruption point, editing phase and repair for re-starts, false starts, slips, PwWs, prolongations.

We examined five types of disfluencies from the aspects of repair processes, those involving real errors (false starts and Ph-errors) and those involving no surface errors but showing interruption and completion of the words (re-starts, PwWs and prolongations) indicating some kind of covert error during speech planning. Trouble signals are not confirmed in the latter cases and the speakers select from the possible simplest solutions in the given context to finish word articulation. The questions arise: (i) Is the place of interruption characteristic of the disfluency in question? (ii) Do durations of the editing phases reflect monitoring and repair processes in these five disfluencies irrespective of the source of the trouble behind them? Our purpose was to show whether the place of interruption and durations of the editing phases would shed light on the specific properties of the analyzed disfluencies.

We had three hypotheses. (1) Halt command at a definite time point in word articulation would reflect the type of disfluency. (2) Editing phase duration is characteristic of disfluency type. (3) There will be a close interrelation between the quantity of articulated segments of abandoned words and the durations of the editing phases.

Methodology

Spontaneous narratives and conversations of 52 Hungarian-speaking subjects (aged between 20 and 40 years, half of them were females) were randomly selected (with the exceptions of age and gender) from the BEA Spontaneous Speech Database of Hungarian (Gósy, 2012). All speakers had normal hearing, and none of them had any speech defects. All of them had a similar socio-economic status.

Speakers were asked to speak about their family, life, hobby, to share their opinion on a specific topic raised by the interviewer, and to summarize two short stories they heard during the interview. In addition, a 3-member conversation about various topics with each participant was subjected to analysis. A total of 27.7 hours of spontaneous speech material was analyzed. The average length of the speech material per speaker was about 30 minutes. Five types of disfluencies were identified with all speakers (their occurrences showed large individual differences). Altogether 1,472 disfluencies were identified in the corpus, and about 28 instances could be found per speaker (also with large differences across subjects). Disfluencies that concerned content words were considered.

All editing phases were manually annotated in the waveform and spectrogram displays via continuous listening to the words in Praat (Boersma & Weenink, 2014). Durations of the editing phases were taken by measuring them between the interruption point (the last segment of the reparandum) and the onset of the continuation (the first segment of the repair/continuation). In prolongations, 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 neighborhood context, as appropriate. Prolongations were identified by the author based on her native language competence. To test statistical significance, the GLMM method, and the Kruskal-Wallis test were used (as appropriate) using SPSS software 21.0. The confidence level was set at the conventional 95%.

Results

Prolongations turned out to be the most frequent type (29%), followed by PwWs (28.5%), Ph-errors (17.3%), re-starts (12.8%), and finally false starts (12.4%). PwWs had three subtypes according to the interruption point. Out of all PwWs 59.5% occurred between stems and suffixes, 21.4% within the stem, while 19.1% of them were compounds with the pause

between the two constituents (for example, *gyerekek* /pause/ *ről* 'children /pause/ about'; *ver* /pause/ *senyez* 'com /pause/ pete'; *csecsemő* /pause/ *otthon* 'infant /pause/ nursing home', respectively). Real errors occurred in one third of all disfluencies. The frequency of the five disfluencies turned out to be 0.88 incidents per minute (prolongations: 0.26, re-starts: 0.11, Ph-errors: 0.15, PwWs: 0.25, false starts: 0.10 per minute) with significant differences in the occurrences of various types of disfluencies (*Chi-Square* = 252.813, $p = 0.001$). Halt command was received after articulating various numbers of segments from 1 to 9 (see Figure 3).

The occurrence of all disfluencies according to pronounced number of segments shows a decrease (27.1%, 18.7%, 18.0%, 17.4%, 13.9%, 10.7%, 5.6%, 4.0%, 2.7%, respectively). The distribution of the incidents across the quantities of abandoned words are characteristic of the disfluency types (Figure 2). The larger the pronounced part of the word the less frequent the occurrence of the disfluency.

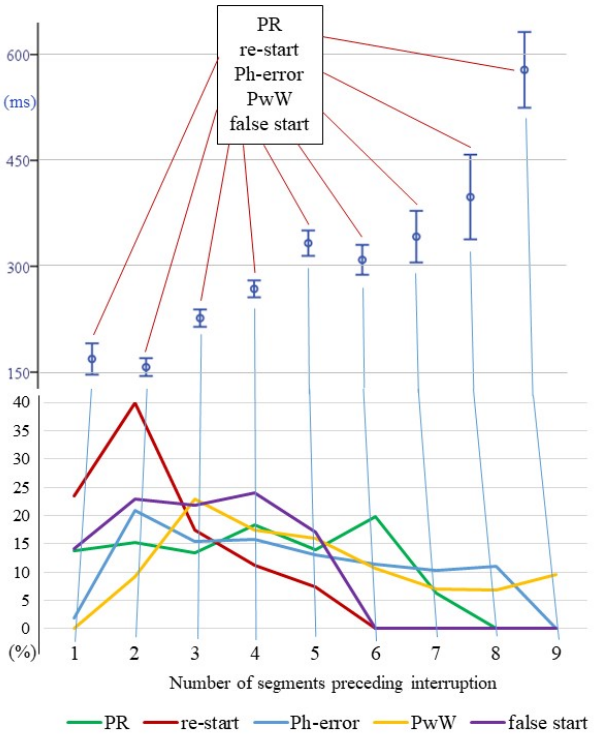


Figure 2. Occurrences of disfluency types depending on the pronounced segments prior to interruption.

No interruptions in words were found in 13.5% of all incidents of re-starts, false starts and PwWs (there were no significant differences among them). Durations of editing phases were the shortest in re-starts (mean/SD = 142/103 ms) and Ph-errors (161/107 ms), longer in prolongations (274/79 ms) and false starts (236/153 ms), and the longest in PwWs (376/168 ms). The editing phase durations of the PwW subtypes show large differences (between stem and suffix: 424/179 ms; within the stem it is 278/115 ms; and before the second word of a compound it is 333/134 ms). Significant differences were confirmed in the editing phase durations among various types of disfluencies ($F(4, 1471) = 169.551$, $p = 0.001$) while pairwise tests revealed that there were no significant differences between re-starts and Ph-errors or between PwW subtypes where a pause occurred in stems and in compounds.

Durations show an increase as the number of pronounced segments increases before halting with the only exception of prolongations ($F(8, 1471) = 90.983$, $p = 0.001$). However, no statistically confirmed durational differences were found in some cases, between 1 and 2, 5 and 6, 5 and 7, 6 and 7, and 7 and 8 pronounced segments. In general, the more segments are pronounced prior to halt the longer the editing phases (Figure 3).

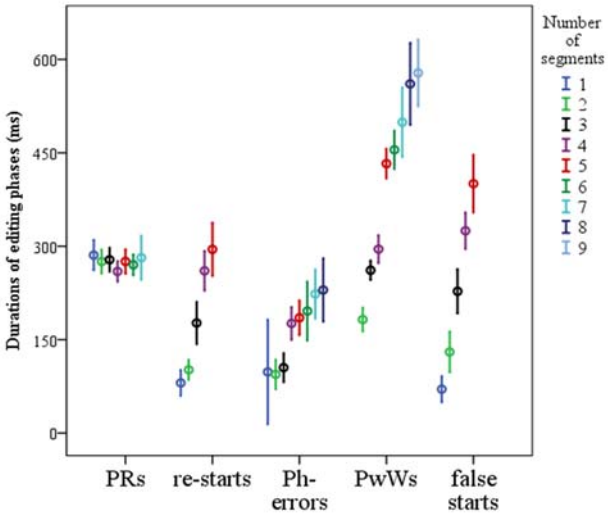


Figure 3. Durations of editing phases depending on both the disfluency type and the number of pronounced segments prior to interruption.

Up to three articulated segments before halting result in a relatively short (below 300 ms) editing phase. From the fourth segment, editing phase durations show a sharp increase, particularly in PwWs. The distribution of editing phase durations seems to be characteristic of the disfluency type.

Discussion

This study focused on editing phase durations in five types of disfluencies using a large Hungarian spontaneous speech database. Findings showed that both the interruption points in the words and the durations of the editing phases and their interrelations are characteristic of the analyzed disfluencies.

We assumed that halt command in word articulation would reflect the type of disfluency. Data supported the hypothesis. The point when articulation is interrupted indicates the speed of action decision after the monitor signals some trouble. This speed seems to be the fastest in restarts, followed by Ph-errors (Figure 2). The difference between them after the third segments before interruption is explained by the different nature of the problems. The number of articulated segments in re-starts decreases after the third segment: the monitor works quickly and accurately. The interruption points of PwW phenomena in stems and in compounds are similar to those of re-starts. The interruption points show some "delay" in the cases of false starts, the decrease can be seen after the fifth segment of the articulated word. It seems that it is easier to approve of a word start than to accept it to be false. Prolongations are distributed in the segments of word articulation similarly to the occurrences of incidents of the PwW subtype where a pause occurs between stem and suffix. The various intensity levels of the trouble signal and the speed of operating monitor are supposed to trigger the speaker's strategy to solve the problem.

Durations of editing phases support our observations based on interruption points. Repair of re-starts is relatively fast; all other disfluency phenomena require longer time. Repairs of a misarticulated segment can be done faster than in false starts where the speaker is uncertain even about lemma selection. They are in a way in connection with the undeniable advantage that the intended word is at hand, albeit erroneously, in Ph-errors. Pause within a word stem is significantly longer than pauses in re-starts: the speaker's strategy is different in the two cases. The temporal difference can be explained by both the diverse monitoring outcome and problem solution. Pauses in compounds are longer than those occurring in stems by 56 ms (the difference is insignificant, though), and can be explained by the semantically narrower bunch of lexemes in the latter case as opposed to the former type. In the retrieval of the second word of a compound (PwW), the lemma level is reached but something prevents speech planning to step forward to lexeme level. We suggest that the relatively long editing phases before suffixes in PwW phenomena refer to monitoring not only the necessary suffixes but the whole grammatical structure of the context. Prolongations may reflect the speaker's trouble in lexical access or with competing (activated) lemmas but there are many other possible reasons resulting in prolongations. Speakers can be in need of extra time to select the appropriate suffixes, monitor what has been said, look for the next word or concept.

Based on measured values, a temporal hierarchy can be suggested starting with re-starts that have the shortest editing phases followed by Ph-errors, false starts and prolongations, and finally PwWs. To make the pattern a bit complicated, editing phases of PwWs in word stems are similar to those of false starts and prolongations, while those occurring before suffixes and after the first content word in a compound have the longest lingering periods. We need to emphasize that behind these findings there can be various speech planning and execution difficulties that may or may not concern lexical retrieval. We assumed that there would be a close interaction between the quantity of articulated segments of abandoned words and the durations of the editing phases. Data confirmed that faster realization of error or of any trouble in the speech planning mechanism leads to shorter editing phase.

Conclusion

We can conclude that (i) both the number of pronounced segments of a word and the corresponding editing phases are characteristic of a specific disfluency type, and (ii) speakers select a strategy to overcome their speech planning difficulties most economically considering the actual word retrieval, grammatical formulation of thought, and some other factors that seem to be decisive in that specific context.

References

Blackmer, E. & J. Mitton. 1991. Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition* 39(3): 173–194. https://doi.org/10.1016/0010-0277(91)90052-6

Boersma, P. & D. Weenink. 2014. Doing phonetics by computer (version 5.4). http://www.praat.org/ (accessed 22 October 2014).

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

Huettig, F. & R. J. Hartsuiker. 2010. Listening to ourself is like listening to others: External, but not internal, verbal self-monitoring is based on speech perception. *Language and Cognitive Processes* 25(3): 347–374. https://doi.org/10.1080/01690960903046926

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

Nooteboom, S. G. & H. Quené. 2008. Self-monitoring and feedback: A new attempt to find the main cause of lexical bias in phonological speech errors. *Journal of Memory and Language* 58(3): 837–861. https://doi.org/10.1016/j.jml.2007.05.003

Postma, A. 2000. Detection of errors during speech production: a review of speech monitoring models. https://doi.org/10.1016/S0010-0277(00)00090-1

Five pieces of evidence suggesting large lookahead in spontaneous monologue

Kikuo Maekawa
Division of Spoken Language, The National Institute for Japanese Language and Linguistics, Tokyo, Japan

Abstract

There is considerable disagreement among the researchers of speech production with respect to the range of lookahead or pre-planning. In this paper, five pieces of evidence suggesting the presence of relatively large lookahead in spontaneous monologues are presented, based on the analyses of the Corpus of Spontaneous Japanese. This evidence consistently suggests that the range of a lookahead is six to seven accentual phrases long, which corresponds on average to 3–4 seconds in the time domain.

Introduction

There is wide consensus among researchers of speech that human speech production involves some sort of 'lookahead' or 'pre-planning,' a process whereby a preverbal message is transformed into a verbal one prior to phonological encoding. There is, however, a lack of consensus with regard to the size of the lookahead; some say that the lookahead is, in principle, a single word (Levelt, 1988), while others suggest its domain is as wide as an intonation phrase (Keating & Shattuck-Hufnagel, 2002).

Part of the discrepancy stems from the confusion of linguistic levels. It is not surprising if the amount of lookahead observed at one level is considerably different from the lookahead observed in a different level; Levelt's study, cited above, is primarily concerned with the level of the lexicon, while Keating and Shattuck-Hufnagel (2002) are concerned with prosodic structure, which often spans a domain much wider than a word.

Another factor in this discrepancy, is the insufficiency in the quantitative analysis of spontaneous speech. It goes without saying, that research on lookahead needs to be based on the analysis of spontaneous speech. But, as Nooteboom (1995) points out, such a study is difficult to conduct, primarily due to the limitation in the size of spontaneous data available for analysis.

In this paper, evidence suggesting the presence of unexpectedly large lookahead in spontaneous monologue in Japanese will be presented; but before that, the results of previous studies that dealt with the issue of lookahead in spontaneous Japanese, either directly or indirectly, will be reviewed briefly.

Watanabe (2009) reported a positive correlation between the length of silent pauses and the grammatical complexity of the upcoming clauses; The more complex the upcoming clauses, the longer the pause.

Similarly, Koiso and Den (2015) found a direct causal relationship between four types of speech disfluencies and the complexity of the constituents that followed weak clause boundaries.

Other studies observed f0 and the length of utterance; they revealed an anticipatory f0 rise at the beginning of utterance (Ishimoto, Enomoto & Iida, 2011; Koiso & Ishimoto, 2012; and Maekawa, 2017).

Data

The analyses reported below are all concerned with the Core part of the Corpus of Spontaneous Japanese, or CSJ (Maekawa, 2003). The CSJ-Core is a speech corpus of 500,000 words (44 hours) spoken by 201 speakers of standard Japanese. It consists mainly of recordings of spontaneous monologues such as academic presentations and simulated public speaking by paid subjects of balanced in ages and genders. The CSJ was used in most of the studies reviewed in the previous section.

The CSJ-Core is annotated richly with respect to word segmentation, POS classification, clause boundary labeling, segmental and prosodic labeling by means of the X-JToBI scheme (Maekawa et al., 2002), and the bunsetsu-based dependency structure, among others.

The relational database (RDB) version of the corpus (Koiso et al., 2014) was used in the following analyses.

Analysis

Anticipatory shortening in AP

Anticipatory shortening implies a shortening of stressed syllable duration as a function of the number of other syllables within a word (Nooteboom & Slis, 1972; Bishop & Kim, 2018). This effect has not been reported for Japanese, which is a mora-timed language (Mora is the unit of phonological length in Japanese. In most, but not all cases, it coincides with the syllable).

Figure 1 shows the relationship between the length of accentual phrases (AP), as measured by the number

https://doi.org/10.21862/diss-09-003-maekawa

of constituent mora (abscissa), and the mean speaking rate (ordinate, unit is [mora/sec]). AP is the basic constituent unit of Japanese intonation, which is marked by f0 rise in the beginning and various boundary pitch movements in the end (see Pierrehumbert & Beckman, 1988).

The mean speaking rate of each AP length in the abscissa, is represented by a red circle with the standard error. The blue line is the LOESS non-parametric regression curve, and the shaded area around the curve is the 95% confidence interval. The LOESS curve was computed using the `ggplot2` library (Ver. 3.1.0) of the R language (Ver. 3.5.1) with the smoothing parameter (`'span'`) set to 0.9. Figure 1 is the case that pooled all 67,923 APs (excluding APs longer than 16 morae). The same tendency is observed in all AP locations (1st, 2nd, …, Nth) in an utterance. This figure shows clearly the presence of lookahead at the level of AP.

AP-internal anticipatory f0 rise

Figure 2 shows AP-internal anticipatory f0 rise. The method of presentation is basically the same as in Figure 1. The abscissa stands for the length of AP and the ordinate stands for the *z*-normalized logarithm of f0. The ordinate value plotted here is the difference between the AP-initial low tone (%L), and the phrasal high tone (H−). See Pierrehumbert and Beckman (1988) for the tonal structure of AP in Japanese. The longer the AP, the larger the phrase-initial f0 rise.

Utterance-internal anticipatory f0 rise

Figure 3 shows the anticipatory f0 rise as observed in the beginning of utterance. The abscissa and ordinates stand, respectively, for the length of utterance measured in terms of the number of constituent APs, and the *z*-normalized logarithm of
the mean f0 of the first AP of utterance. There is a clear positive correlation between the utterance length and the mean f0 of the first AP, in the range between 2–6 APs in the abscissa, before the curve reaches the plateau. Note that utterances that consist of a single AP are omitted from the analysis. Note also, that unlike the Figures 1 and 2 that dealt with the lookahead within an AP, Figure 3 shows the evidence of lookahead in much larger prosodic domain.

Duration of the silent pause

As noted above, Watanabe (2009) found a correlation between the silent pause length and utterance length. Here, similarly, Figure 4 plots the relationship between the length of the upcoming utterance (number of constituent AP, abscissa), and the duration ([sec.]) of the silent pause preceding the utterance. Note that exceedingly long silent pauses (those longer than 3 seconds) were omitted from the analysis, because long silent pauses are often caused by troubles that are external to speech, for example,

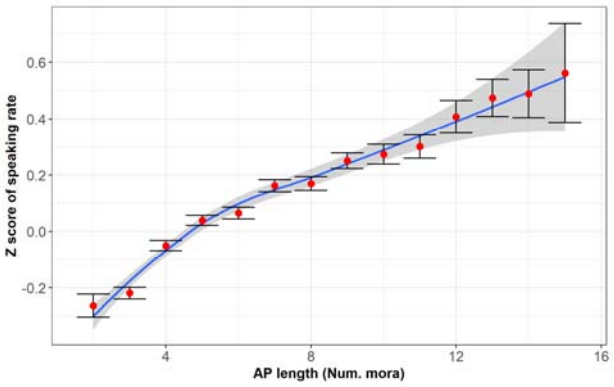


Figure 1. Anticipatory shortening within an AP.

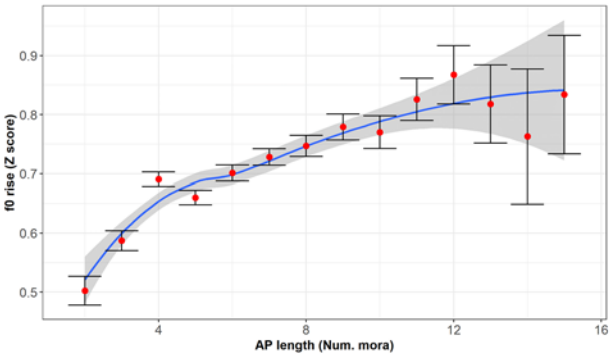


Figure 2. Anticipatory f0 rise within an AP.

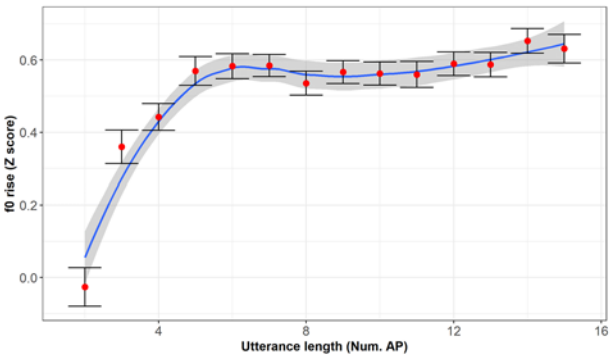


Figure 3. Anticipatory f0 rise within an utterance.

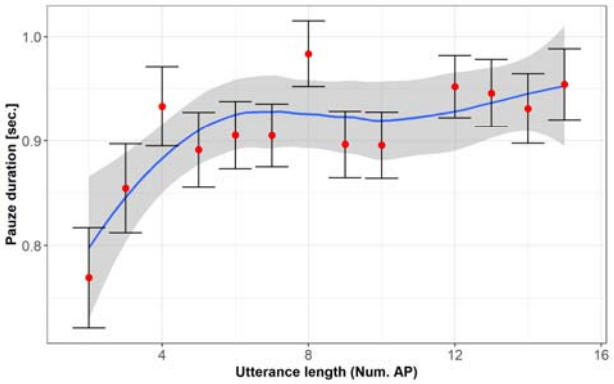


Figure 4. The length of upcoming utterance and the duration [sec.] of utterance-initial silent pause.

difficulties in handling presentation material such as the overhead projector and microphone. Here, the pause duration increases within the range of 1–5 APs, before it reaches the plateau. Note also that this figure has a relatively large confidence interval.

Dependency distance and the f0

Lastly, the relation with syntactic complexity is examined using the dependency structure. In traditional Japanese grammar and Japanese natural language processing, syntactic structure is often represented by means of 'bunsetsu', which is a surface syntactic unit that usually consists of one content word followed by one or more function words.

In Figure 5, the upper part shows the bunsetsu dependency relationship in the example of "kinoo Taro-to Hanako-ga Kyoto-made it-ta" ("Yesterday Taro and Hanako went to Kyoto"), where the time adverbial "kinoo" at the beginning of the sentence modifies the last bunsetsu, which is the predicate and the fourth bunsetsu from the time adverbial. On the other hand, the second bunsetsu modifies the third one. In these cases, the dependency distance is counted as 3 and 0 respectively; the general rule is that if a modified bunsetsu is *N* phrases apart from the modifying bunsetsu, the distance is *N*−1.

The lower part of the figure shows the dependency relation among APs. In Japanese, it is often the case that more than one bunsetsu merged into a single AP. In this example, the second and third bunsetsu on the one hand, and the fourth and fifth on the other, are merged. The dependency distance between these APs is counted in the same manner as in the case of bunsetsu (see Figure 5).

Figure 6 shows the relationship between the dependency distance of a given AP and the mean difference in the *z*-normalized logarithm f0, between the AP in question and the phrase that follows immediately.

This figure shows, that the behavior of f0 is different depending on the presence of discontinuity in the dependency; when there isn't discontinuity (i.e. the distance is zero), the f0 difference is negative, implying that the f0 of the following AP is lower than the AP in question; on the other hand, when there is large discontinuity (e.g. the distance is larger than two), the difference is positive, meaning that the following AP has higher f0 than the AP in question.

Note that the f0 rise in the case of discontinuous dependency cannot be explained as a simple resetting of downstep (Pierrehumbert & Beckman, 1988). Here, the f0 difference between the two APs is not positive when the distance is one, while most theories of downstep (Kubozono, 1993 among others) predict that the resetting occurs whenever there is discontinuity. In any case, the correlation between the distance in dependency, and the f0 behavior can be observed in the range between 0 and 5 or 6 in the abscissa. It is equivalent to saying that the maximum range of lookahead is 6 or 7.

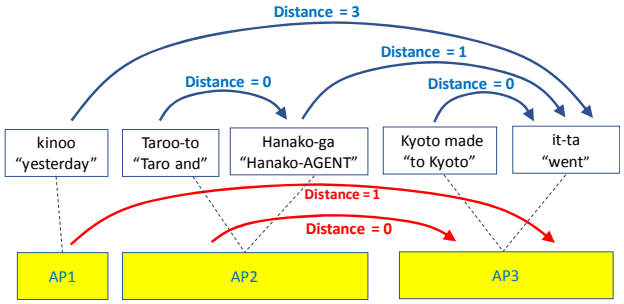


Figure 5. Schematic representation of the dependency distance among bunsetsu and accentual phrases.

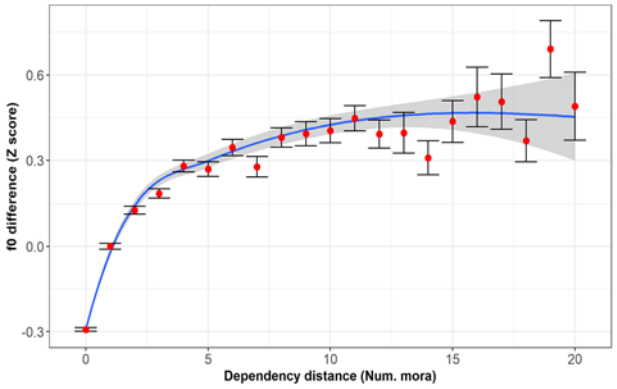


Figure 6. Dependency distance of an AP and the mean f0 difference between the AP and the immediately following AP.

Discussions and conclusion

The results reported in the previous section are the evidence of the presence of lookahead in spontaneous monologue of Japanese. They include evidence for lookahead both inside and outside of AP. Inside an AP, the relationship between the AP length and the mora shortening, or AP-initial f0 rise, was almost linear without any plateau effect; this suggests that there is no clear limit in the size of lookahead within an AP.

On the other hand, there were limits in the size of lookahead observed at the level of utterance; all the figures presented above the curves suggesting the presence of lookahead showed clear plateau effects. More importantly, the maximum size of lookahead estimated independently, coincided almost with the values between 5 and 7 APs.

Given that the mean duration and mean mora length of AP in the CSJ-Core are 0.56 sec and 4.91 morae respectively, the size of lookahead suggested above corresponds to about 2.8–4.0 sec and 25–35

morae. To which linguistic unit, then, does this rather large domain of utterance-level lookahead correspond?

The most obvious possibility is the intonation phrase. In Japanese, an intonation phrase, aka intermediate phrase, is a prosodic domain consisting of several APs, and it is presumed to be the domain of downstep (Pierrehumbert & Beckman, 1988). Several studies concluded that the domain of lookahead in English is the intonation phrase (Keating & Shattuck-Hufnagel, 2002; Bishop & Kim, 2018).

In Japanese, however, this hypothesis is difficult to support. One negative evidence is that, in the case of the CSJ-Core, the mean length of the intonation phrase, which is defined as the prosodic domain as demarcated by the boundary indices (BI) 3 on both edges in the X-JToBI annotation, is much shorter than the size of lookahead estimated in the current study.

In the CSJ-Core, nearly one-third of the intonation phrases coincide with AP (i.e. intonation phrase consists of a single AP), and the mean length of the intonation phrases is about 1.70 AP. Even if we remove the cases of a single AP phrases from the computation, the mean length is no longer than 2.64.

The second negative evidence comes from results in Figure 6. This figure showed that the difference of f0 between the AP in question and the immediately following AP is, on average, close to zero when the dependency distance is 1, and it keeps increasing until it reaches the plateau of about 0.2 at around the distance of 6.

Since a positive value in the f0 distance implies the resetting of downstep (Kubozono, 1993), and thereby the presence of an intonation phrase boundary, the figure suggests that there are many instances where intonation boundaries are included within the domain of lookahead.

At the present stage of this study, it is difficult to conclude exactly what the domain of lookahead corresponds to, but it seems plausible that the domain is larger than the intonation phrase. It might be a type of clause; or it might be that a simple correspondence between the domain of lookahead and linguistic or prosodic structure does not exist. It would not be surprising if the domain of lookahead differs considerably from one speaker to another, reflecting the difference in their working memories.

Acknowledgments

This work is supported by the Kakenhi grants (17H02339 and 19X21641) and the budget of the Center for Corpus Development, NINJAL. The author thanks his colleagues in NINJAL for their comments on earlier version of the paper.

References

Bishop, J. & B. Kim. 2018. Anticipatory shortening: Articulation rate, phrase length, and lookahead in speech production. In: K. Klessa, J. Bachan, A. Wagner, M. Karpiński & D. Śledziński (eds.), *Proceedings of Speech Prosody*, 13–16 June 2018, Poznań, Poland, 13–17. https://doi.org/10.21437/SpeechProsody.2018-48

Ishimoto, Y., M. Enomoto & H. Iida. 2011. Projectability of transition-relevance places using prosodic features in Japanese spontaneous conversation. In: P. Cosi, R. De Mori, G. Di Fabbrizio & R. Pieraccini (eds.), *Proceeding of Interspeech,* 27–31 August, Florence, Italy, 2061–2064.

Keating, P. & S. Shattuck-Hufnagel. 2002. A prosodic view of word form encoding for speech production. *UCLA Working Papers in Phonetics* 101: 112–156.

Koiso, H. & Y. Den. 2015. Causal analysis of acoustic and linguistic factors related to speech planning in Japanese monologs. In: *DiSS 2015, Proceedings of the 7th Workshop on Disfluency in Spontaneous Speech*, 8-9 August 2015, University of Edinburgh, Scotland, UK.

Koiso, H. & Y. Ishimoto. 2012. Nihongo hanashikotoba koopasuo mochiita hatsuwano inritsutekitokuchoono bunseki: Intoneeshonkuwo kirikuchito shite [Prosodic Features of Utterances in the Corpus of Spontaneous Japanese: Intonational Phrase–Based Approach]. In: *Proceedings of the 1st Corpus Japanese Linguistics Workshop*, 56 March 2012, Tokyo, Japan, 167–176.

Koiso, H., Y. Den, K. Nishikawa & K. Maekawa. 2014. Design and development of an RDB version of the Corpus of Spontaneous Japanese. In: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of LREC 2014*, 26 May 2014, Rejkjavik, Iceland, 311–315.

Kubozono, H. 1993. *The Organization of Japanese Prosody*. Tokyo: Kuroshio Publishing.

Levelt, W. J. M. 1988. *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Maekawa, K. 2003. Corpus of Spontaneous Japanese: Its Design and Evaluation. In: *Proceedings of SSPR 2003*, 13–16 April 2003, Tokyo, Japan, 7–12.

Maekawa, K. 2017. A new model of final lowering in spontaneous monologue. In: *Proceedings of Interspeech 2017*, 20–24 August 2017. Stockholm, Sweden, 1233–1237. https://doi.org/10.21437/Interspeech.2017-175

Maekawa, K., H. Kikuchi, Y. Igarashi & J. Venditti. 2002. X-JToBI: An extended J_ToBI for spontaneous speech. In: *Proceedings of ICSLP 2002*, 16–20 September 2002, Denver, CO, 1545–1548.

Nooteboom, S. G. 1995. Limited lookahead in speech production. In: F. Bell-Berti & L. R. Raphael (eds.), *Producing speech: Contemporary issues—for Katherine Safford Harris*, NY: AIP Press, 3–18.

Nooteboom, S. G. & I. H. Slis. 1972. The Phonetic Feature of Vowel Length in Dutch. *Language and Speech*, 15(4): 301–316. https://doi.org/10.1177/002383097201500401

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

Watanabe, M. 2009. *Features and Roles of Filled Pauses in Speech Communication*. Tokyo: Hituzi.

Fill the silence! Basics for modeling hesitation

Simon Betz [1] and Loulou Kosmala [2]
[1]Phonetics and Phonology Workgroup, Bielefeld University, Bielefeld, Germany
[2]PRISMES EA 4398/SeSyLiA, Sorbonne Nouvelle University, Paris, France

Abstract

In order to model hesitations for technical applications such as conversational speech synthesis, it is desirable to understand interactions between individual hesitation markers. In this study, we explore two markers that have been subject to many discussions: silences and fillers. While it is generally acknowledged that fillers occur in two distinct forms, um *and* uh*, it is not agreed on whether these forms systematically influence the length of associated silences. This notion will be investigated on a small dataset of English spontaneous speech data, and the measure of distance between filler and silence will be introduced to the analyses. Results suggest that filler type influences associated silence duration systematically and that silences tend to gravitate towards fillers in utterances, exhibiting systematically lower duration when preceding them. These results provide valuable insights for improving existing hesitation models.*

Introduction

When speakers are engaged in face-to-face interactions, their productions contain frequent hesitations. Hesitation can be defined as "the temporary suspension of flowing speech" (Lickley, 2015: 40). This paper focuses on the distribution and duration of two common hesitation markers: fillers and silences, especially in co-occurrence, i.e. silences that appear in the same utterances as fillers.

Fillers and silences are said to be the most frequent types of hesitations (Shriberg, 1994, Eklund, 2004). Their temporal features have been explored by numerous researchers, and some of them have labelled them as signals of suspension (Clark & Fox Tree, 2002). Clark and Fox Tree's main hypothesis is that fillers signal a speaker's intention to initiate a delay, and that um signals a major delay, while uh signals a minor one. This was in part explained by the fact that more pauses occurred with um than with uh. The presence of a pause can thus play a role in this signaling-a-delay hypothesis. Other studies have looked at the co-occurrence of fillers and silences: Grosjean and Deschamps (1972) found that fillers were often combined with silences both in French and English; Smith and Clark (1993) argued that um was followed by a longer pause than uh because speakers intentionally chose between uh and um to signal their word retrieval difficulties; in a study of pauses in deceptive speech, Benus et al. (2006) found that um was more likely to be followed by a silence than uh, and that silences were longer when they followed turn-initial um. In investigating the cluster of disfluencies, Kosmala and Morgenstern (2017) found two recurrent combinations: filler + silence, and lengthening + filler. Betz and Lopez Gambino (2016) also found that speakers engaged in a description task sometimes produced a filler after an initial silence, which allowed them to buy time before planning the description.

However, the idea of uh and um being consciously chosen by speakers to signal an upcoming delay is questionable. Finlayson and Corley (2012) argued that the fact that fillers tend to precede silences does not necessarily mean that they are intentionally chosen. O'Connell and Kowal (2005) rejected the signaling hypothesis and more specifically the status of uh and um as interjections, and Schegloff (2010: 71) argued that although fillers can be associated with delay, they do not "announce" a delay, but rather "embody" it.

In line with these issues, we further explore the co-occurrence of fillers and silences. Clusters of multiple markers have seldom been the focus of analysis, so our aim is to provide insights about the interplay of hesitation markers in order to model them for technical applications. We focus on two broad topics in this investigation. First, we test if the challenged assumption by Clark and Fox Tree—that silence duration varies as a function of filler type—can be confirmed. Second, we extend the analysis by measuring distance between silences and fillers to test whether fillers can influence duration of silences that are further remote than their direct vicinity.

Corpus and methods

The materials used for this study are taken from the FILM corpus (Kosmala & Morgenstern 2019) which is a collection of recordings between 16 native English speakers (aged 18–23) engaged in face-to-face dyadic interactions in the form of a film interview in familiar settings. The participants knew each other fairly well, and interacted in pairs. The interviewer asked a series of 10 questions about the film to the interviewee, and the latter was asked to answer the questions as spontaneously as possible.

https://doi.org/10.21862/diss-09-004-betz-kosm

The total duration of the corpus is approximately 71 minutes.

We investigated the co-occurrence of fillers and silences in the data. Following Clark and Fox Tree (2002), we distinguished two types of fillers, uh and um. These fillers differ on the phonetic surface in being either a centralized vowel (uh) or a centralized vowel with a nasal (um). As explained earlier, according to Clark and Fox Tree (2002), these two types are mutually exclusive and denote either a minor or major delay, which can be quantified by measuring adjacent silences. For this investigation, we thus measured the duration of fillers and associated silences (i.e. silences in the same utterance, either preceding or following), using the ELAN software. We were only interested in co-occurring hesitations so we only selected utterances that contained both fillers and silences, yielding 722 silences and 303 fillers in total. For silences, we measured the distance in words from the fillers. 0 denotes the first position after a filler, values > 0 subsequent positions. -1 denotes the last position before the filler, values < -1 greater distance before a filler.

We first aim to test the hypotheses stated by Clark and Fox Tree (2002):

- Is the duration of silences associated with um higher than those associated with uh?
- Does the duration of the filler correlate positively with the duration of the associated silence?

Furthermore, we explored the notion of distance:

- What is the average distance of silences in utterances where they co-occur with fillers?
- Does the distance between silence and filler influence silence duration?

Results

Silence duration after um and uh

The duration of silences is on average 155 ms higher when the silence occurs in an utterance with a um-type filler as opposed to uh-type filler. We fitted a linear mixed effects model with silence duration as the dependent variable and filler type as the independent variable. We included as random effects random slopes for speakers, random slopes for distance between silence and filler, and random slopes for position of the filler within the utterance. The difference is significant: $p = 0.0084$, $t = 2.661$, $DF = 210.57$, $SE = 56.56$. The speakers showed great variability. Model comparisons using analyses of variance between the full model and the reduced model without random slopes for speakers yielded significant results ($p = 0.019$). This is likely due to the fact that the amount of fillers produced varies strongly per speaker, which has been attested for this dataset (Kosmala & Morgenstern, 2019) and has frequently been observed with other data as well (e.g. Betz & Lopez Gambino, 2016). We conducted exploratory post-hoc tests to see if filler rate per speaker correlates with silence or filler duration, but found no such interaction.

This study is based on a small dataset of interview-style interaction, in which we expect a lot of turn-initial fillers, but the model comparisons suggested no influence of position on the results. We conducted additional *t*-tests on the mean duration of utterance-medial and utterance-initial fillers and associated silences. The general idea that silences are longer when they co-occur with um-type fillers was confirmed for medial position only, but not for initial position, where the same tendency was observed, but failed to reach significance ($p = 0.044$, $DF = 56.4$ $t = 2.06$ for medial position; $p = 0.21$, $DF = 42.9$, $t = 1.26$ for initial position). Additionally, silences associated with fillers of both types in utterance-initial position were longer than those in medial position, but not significantly (for um: $p = 0.14$, $DF = 83$, $t = 1.5$; for uh: $p = 0.12$, $DF = 45.21$, $t = 1.6$). This difference might be clearer when analyzed on a bigger dataset, as it would be conceivable that turn-initial, planning-related hesitations span a significantly longer time.

Correlation of filler and silence duration

Clark and Fox Tree (2002) split the two types of fillers into prolonged and not prolonged fillers, yielding four types (um, u:m, uh, u:h). They found that prolonged fillers were associated with longer silences. In our data, the types were not divided a priori between prolonged and not prolonged; we rather fit linear regression models to see if duration of a filler correlated with the duration of the associated silence.

Our data contains 303 fillers, 83 of which appeared with no silences. We conducted this analysis for both uh and um, confirming the findings of Clark and Fox Tree (2002): longer uh and um were associated with longer silences in the utterance ($p = 0.017$, $t = 2.4$, $SE = 0.04$ for um; $p = 0.009$, $t = 2.7$, $SE = 0.07$ for uh). This finding, however, has to be taken with caution. When considering silences immediately following or preceding the filler, only the duration of silences preceding um correlated with the duration of the filler. This might be an artifact of the small size of our dataset and is up for future research to verify.

Distance between silences and fillers

As Figure 1 shows, the longest silences are directly adjacent to fillers, either preceding (-1) or following

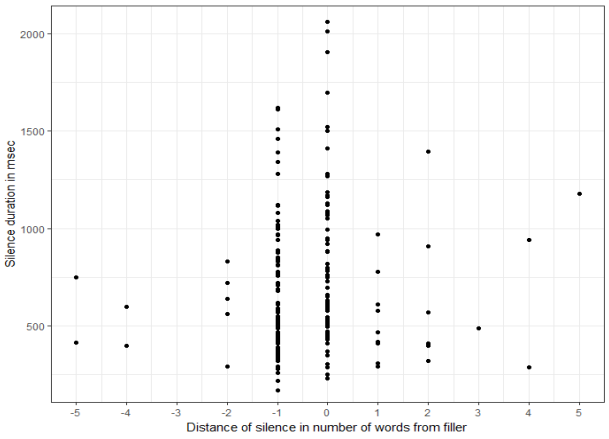


Figure 1. Silence duration and distance from fillers.

(0). The mean distance is lower for um (−0.6) than for uh (−0.5), but as a *t*-test reveals, this difference is not significant ($p = 0.7$, $DF = 83.7$, $t = -0.375$). However, the distance seems to influence the duration of silences. As can further be seen, silences before fillers (−1) are shorter than following fillers (0). This difference is significant and holds true for both um and uh type fillers ($p = 0.013$, $DF = 78.7$, $t = 2.5$ for um; $p = 0.017$, $DF = 40.4$ $t = 2.5$ for uh).

Discussion

Filler types

The claim that fillers are an intentional signal with a word status has been hotly debated in the research community. This study on a few hundred instances of hesitation clusters does not claim to be a tie-breaker for this discussion. However, our results point to a clear direction, comparable to those by Clark and Fox Tree (2002), differing from those by O'Connell and Kowal (2005). Silence duration does vary systematically depending on filler type, lending support to the very notion that there are indeed two distinct types of fillers.

Inter-speaker variability

Speakers produced on average 38 fillers, but as observed in an earlier study conducted on the same data by Kosmala and Morgenstern (2019), there is great inter-speaker variability in filler usage, a phenomenon which has often been observed in disfluency research. While every speaker has their own preference with regard to hesitation patterns, there seems to be no systematic influence on the variables tested. For this study it is sufficient to observe that the general tendencies observed in earlier studies hold true. But if these findings were put to practical application, such as speech synthesis, a model of one particular speaker might lead to totally different results than a model constructed on the basis of mean values from a pool of speakers.

Speaking style influence

The results may additionally be strongly influenced by the communication settings and the speaking style. Clark and Fox Tree (2002) used spontaneous face-to-face conversations, whereas O'Connell and Kowal (2005) used speech data from a trained and educated speaker being interviewed by media experts. Our analyses were conducted on interview-style data, but spontaneous nonetheless, which might be a reason for the closeness to Clark and Fox Tree's results.

Standard maximum silence

The analyses on distance revealed some further insights. Most silences, when associated with fillers, occurred in direct vicinity to them. There was a significant difference in silence duration preceding and following fillers, which requires some future attention. If this is robust, it could lend support to the notion of Standard Maximum Silence (Jefferson, 1989): there is an upper threshold for silences in conversation, and when it is exceeded, either speaker will contribute to bridge the silence. This follows the initial notion dating back from the beginning of disfluency research in the 1950s that being silent for too long puts the speaker at risk to lose the conversational floor (Maclay & Osgood, 1959). It would make sense then, to have shorter silences before fillers than after, because once the speaker has produced the filler, the listener has already been provided with a cue that conversation might continue, so there is more pressure to fill the silence when no filler has occurred yet.

Application and outlook

One practical application of the results obtained here is the extension of the hesitation insertion model for speech synthesis, which has been prototypically tested in Betz et al. (2018), which did not yet take into account the structural interplay of silences and fillers. Furthermore, the hesitation model by Betz et al. (2018) is centered on lengthening, which provides an elegant entry point for a synthetic hesitation interval, and reflects human speech production by making use of the pre-planned, but not-yet-uttered words in the articulatory buffer (Levelt, 1989). This approach receives support by the confirmed notion of longer silences after um-type fillers: the presence of a nasal sound makes this type of filler a better candidate to smoothly initiate a hesitation interval by lengthening compared to the uh-type fillers (for hesitation lengthening distribution over phone types, cf. Betz, Wagner & Voße, 2016).

For future work on these matters, it is desirable to extend the analyses started here to a dataset with phonemic annotation, so that lengthening can be

included as a third hesitation marker which might frequently cluster with silences and fillers. The hypothesis would be that hesitation lengthening clustered with fillers would presumably be associated with um-type fillers which denote a longer delay. Betz and Wagner (2016) observed that phones preceding fillers undergo the same lengthening processes as phones preceding intonation phrase boundaries. This has been explained by the presence of fillers introducing an additional intonation phrase boundary at positions not predicted by syntax, which causes the typical phrase-final lengthening to occur. However, in that study, fillers were not distinguished into uh and um types.

Conclusion

Hesitation markers occur in speech both in standalone form and in clusters. While clusters are comparably rare, it is still desirable to be able to model them adequately. In Betz, Wagner and Schlangen (2015), it was found that the more hesitation markers were included in the same synthetic utterance, the worse user ratings got. This might well be due to the fact that there are certain syntactic rules which govern how hesitation markers have to be combined. This study was a first step, investigating the much-discussed interplay of silences and fillers, for future work it is desirable to extend these analyses to include the third prototypical hesitation marker, lengthening, in order to get a full picture of the mechanisms behind hesitation clusters.

References

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

Betz, S., B. Carlmeyer, P. Wagner & B. Wrede. 2018. Interactive Hesitation Synthesis: Modelling and Evaluation. *Multimodal Technologies and Interaction* 2(1): 9. https://doi.org/10.3390/mti2010009

Betz, S. & S. López Gambino. 2016. Are We All Disfluent in Our Own Special Way and Should Dialogue Systems Also Be? In: O. Jokisch (ed.), *Studientexte zur Sprachkommunikation 81*. Dresden: TUD Press, 168–174.

Betz, S., & P. Wagner. 2016. Disfluent Lengthening in Spontaneous Speech. In: O. Jokisch (ed.), *Studientexte zur Sprachkommunikation 81*. Dresden: TUD Press, 135–144.

Betz, S., P. Wagner & D. Schlangen. 2015. Micro-Structure of Disfluencies: Basics for Conversational Speech Synthesis. In: *Proceedings of INTERSPEECH*, Dresden, Germany, 2222–2226.

Betz, S., P. Wagner & J. Voße. 2016. Deriving a Strategy for Synthesizing Lengthening Disfluencies Based on Spontaneous Conversational Speech Data. In: C. Draxler & F. Kleber (eds.), *Tagungsband Der 12. Tagung Phonetik Und Phonologie Im Deutschsprachigen Raum*, Munich, Germany, 19–22.

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

Eklund, R. 2004. *Disfluency in Swedish human-human and human-machine travel booking dialogues*. Ph.D. dissertation, Linköping University.

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

Grosjean, F., & A. Deschamps. 1972. Analyse Des Variables Temporelles Du Français Spontané. *Phonetica* 26(3): 129–56. https://doi.org/10.1159/000259407

Jefferson, G. 1989. Preliminary Notes on a Possible Metric Which Provides for a 'Standard Maximum' Silences of Approximately One Second in Conversation. In: D. Roger & P. Bull (eds.), *Conversation: An Interdisciplinary Perspective*, Clevedon: Multinlingual Matters, 166–96.

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

Kosmala, L. & A. Morgenstern. 2019. Should 'uh' and 'um' Be Categorized as Markers of Disfluency? The Use of Fillers in a Challenging Conversational Context. In: L. Degand, G. Gilquin, L. Meurant & A. C. Simon (eds.): *Fluency and Disfluency across Languages and Language Varieties*, Louvain-la-Neuve: Presses Universitaires de Louvain, 67–90.

Levelt, W. J. 1989. *Speaking*. Cambridge, MA: MIT Press.

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

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

O'Connell, D. C. & S. Kowal. 2005. Uh and Um Revisited: Are They Interjections for Signaling Delay? *Journal of Psycholinguistic Research* 34(6), 555–76. https://doi.org/10.1007/s10936-005-9164-3

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

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

Smith, V. L. & H. H. Clark. 1993. On the Course of Answering Questions. *Journal of Memory and Language* 32(1): 25–38 https://doi.org/10.1006/jmla.1993.1002.

Variation in the choice of filled pause: A language change, or a variation in meaning?

Hong Zhang
Department of Linguistics, University of Pennsylvania, Philadelphia, PA, USA

Abstract

The role of filled pauses in message structuring is a heavily debated question, but the result is still somewhat inconclusive. In this study, I consider this question jointly with sociolinguistic factors that have been thought to affect the choice of filled pause in American English. The results suggest that the use of uh *is subject to higher variability across not only age groups, but also conversation topics and interlocutors. A latent semantic analysis found consistent difference between two forms of filled pause and silent pauses of varying duration in the primary latent dimension, but similarity between short silent pause and* uh*, as well as long silent pause and* um *in the second dimension. Therefore, the functional difference between* um *and* uh *should be acknowledged, and the observed change in their relative popularity is potentially related to their different meaning or function in the discourse.*

Introduction

Two forms of filled pause in American English, *um* and *uh*, have long been the focus of studies on speech disfluencies. In addition to understanding their relation to the cognitive process of speech production (e.g. Goldman-Eisler, 1968; Rochester, 1973; Levelt, 1983), debates have also been centered around whether the two fillers play different roles in structuring the message. Clark and Fox Tree (2002) argue that the fillers are different words which carry distinctive discourse meanings, citing evidence from the different delay after the fillers and the different role in facilitating word recognition. However, a later study (O'Connell & Kowal, 2005) failed to find the functional difference through analyzing media interviews of Hillary Clinton. In a review paper by Corley and Stewart (2008), they concluded that there is not clear evidence that speakers have intentional control over the choice of fillers.

More recently, correlations between sociolinguistic factors, such as age, gender and socioeconomic status of the speaker, and variation in the choice of filler forms have been thoroughly examined both in English (Acton, 2011; Tottie, 2011; Fruehwald, 2016) and several other Germanic languages (Wieling et al., 2016). Among the sociolinguistic factors, an interesting observation is that the relative frequency of *um* over *uh* is higher in younger speakers, where female speakers seem to lead the trend. Therefore, it has been proposed that the choice of filled pause is in fact a sociolinguistic variable and a language change in progress in the Germanic family.

However, two potential covariates which in principle partially explain the variation in the choice of filled pause: the topic of conversations and the accommodation between interlocutors, have not been explicitly and systematically discussed in the sociolinguistic literature. In this study, I show that conversation topic and the effect of interlocutor are able to reveal a possible functional difference between the two filled pauses, and this difference may also point to valid explanations of the proposed change in progress. However, this difference may or may not be the kind as proposed in Clark and Fox Tree (2002), and should only be understood in relation with other potential influencing variables.

Data and method

The corpus

The data used for this study is from the Fisher Corpus of American English (Cieri, Miller & Walker, 2004). This corpus contains 16,454 10-minute telephone conversations recorded in separate channels, totaling 2,742 hours of speech. Conversations were guided by 40 topics. A subset of 9,471 one-sided speech from 3,157 native speakers of American English are selected to form the analysis sample, representing about 790 hours of spontaneous conversations. The selected speakers all contributed in exactly three conversations.

Data pre-processing

The analysis sample is forced aligned with Penn Forced Aligner (Yuan & Liberman, 2008). Each conversation was first segmented into turns based on the time stamps and speaker identification number provided in the transcription before passing to the forced aligner. Back channel talking and floor holding were also removed by filtering out turns shorter than four words and containing only filler words. Durations were obtained from forced alignment results.

https://doi.org/10.21862/diss-09-005-zhang

Sociolinguistic factors

Age and gender

Age and gender's effect on the alternation between two forms of filled pause is most profoundly discussed in sociolinguistic literature. The most noticeable change, according to Fruehwald (2016) and Wieling et al. (2016), is that *um* is trading frequency with *uh* among younger speakers. This reported trend serves as the main evidence for the "language change in progress" argument (Labov, 1994).

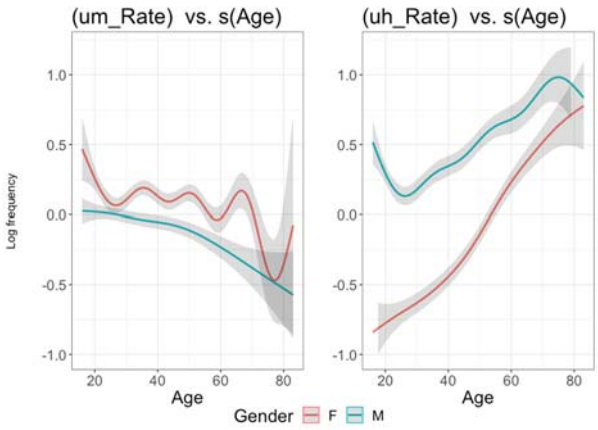


Figure 1. Age and gender effect on the frequency of filled pause by form.

Figure 1 plots the per-speaker frequency of the two fillers as the function of age, grouped by gender. Regression lines are fitted through a quasi-Poisson regression with per-speaker as the response, and shaded areas represent 95% confidence bands. The figure clearly shows that the change in frequency is mainly driven by the change in *uh*. On the other hand, the frequency of *um* is largely stable for female speakers between 20 and 60 years of age, and only a slight decrease can be found among male speakers. Inspecting the effect of age on the overall per speaker frequency of filled pause, as shown in Figure 2, it is observed that there is an overall increase in the frequency as people getting older. Male speakers consistently have higher frequency than female speakers, just as in the case of *uh*. Therefore *um* appears to be time-invariant, while *uh* displays more change as a function of age. Therefore the term 'trading frequency' may be an exaggeration.

Individual variation

The individual variation of *um*/*uh* choice can be effectively visualized through the 2-D density plot with the frequency of *um* and *uh* as two dimensions, as plotted in Figure 3.

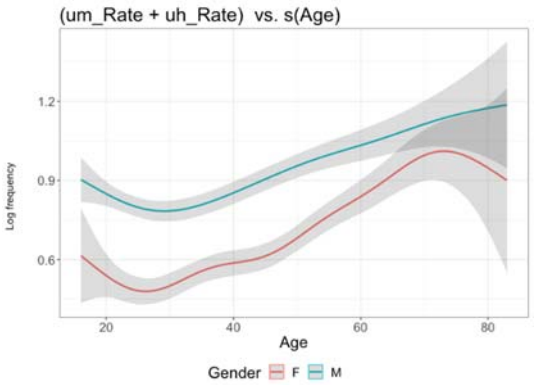


Figure 2. Age and gender effect on the overall frequency of filled pause.

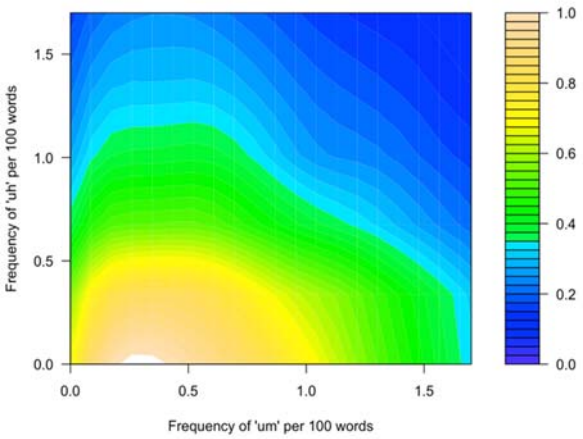


Figure 3. Individual variation in the choice of filled pause plotted as per speaker frequency of um *and* uh.

Two noticeable patterns can be observed from Figure 3. First, there is an inverse correlation between the frequency of *um* and *uh*, but this relation only holds for speakers who use more *um*. Second, the highest density is along the *um* axis, meaning that there are more predominantly *um* users. Thus for a given speaker, there does seem to be a preference for one filler over the other, and *um* appears to be the more popular choice.

Influence of topic and interlocutor

The effect of topic on individual choice of fillers can be visualized on the same dimensions shown in Figure 3. The median frequencies across topics are plotted in Figure 4. It is obvious that greater variation can be found along the *uh* frequency dimension. This pattern offers another piece of evidence that *uh* has greater variability compared to *um* across different conversation topics. Combining the observations from Figure 3, one plausible hypothesis is that for many people, *um* is the default choice of filled pause for the need to insert a filler, while *uh* comes in more involuntarily as a response to other challenges in

speech production (i.e. they have different meanings).

This and other similar hypotheses can be supported by two further observations: The patterns of between-topic and between-speaker correlations of the frequency of fillers. Figure 5 is the quantile plot of between-topic correlations of *um* and *uh* frequencies. The pairwise correlations between topics are not only higher for *um* for about half of the topic pairs, the values of correlation coefficients are also almost consistently greater than 0.9. On the contrary, about 40% of the correlations are less than 0.9 for *uh* frequency, and the lowest correlation is lower than 0.6. Considering the inventory of topics in Fisher, the variation among *uh* frequencies, as well as the contrast between *um* and *uh*, is non-trivial. To give an example, the topics provided to speakers ranged from Iraqi War to Food and Personal Hobbies. The split between more serious social or political topics, and topics on more casual or hypothetical situations is about 3 to 8. More similarities in terms of the nature of content across conversations is expected among the more casual topics. The cumulative density function of *uh* frequency therefore roughly aligns with this topic split.

The between-speaker correlations are reported in Table 1. Here the correlations are calculated from a subset of the analysis sample in which both sides of the conversation appeared in the pool of speakers. Three conditions are separated to account for the potential gender effect on accommodation. This results in a subsample consisting of 685 male-male conversations, 885 female-female conversations and 675 male-female conversations.

Table 1. Correlation of filled pause frequency in conversation between speakers controlled for gender (:p < 0.05).*

	um	*uh*
Male-male conversation	0.113*	0.368*
Female-female conversation	0.103*	0.205*
Male-female conversation	0.056	0.192*

In all conditions, higher correlations are found in the frequency of *uh*. The smallest difference is in the case of female-female conversations, where the magnitude is still two times higher. Although the absolute between-speaker correlations are at most moderate, the clear difference between the two fillers nevertheless suggests that *uh* is more likely to be accommodated by the interlocutor, especially when both sides are males. This is consistent with earlier observations that *uh* is subject to more variation compared to *um*.

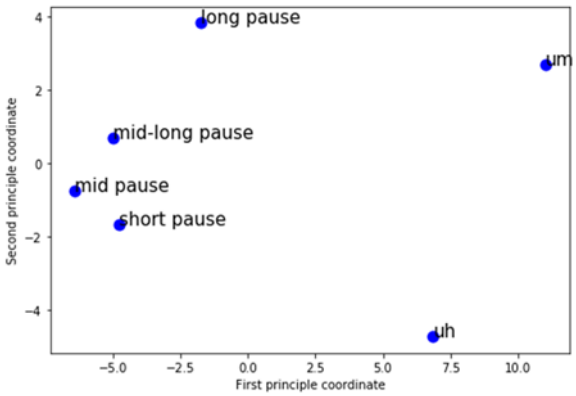


Figure 6. LSA analysis of silent and filled pauses projected to the 2D space using MDS.

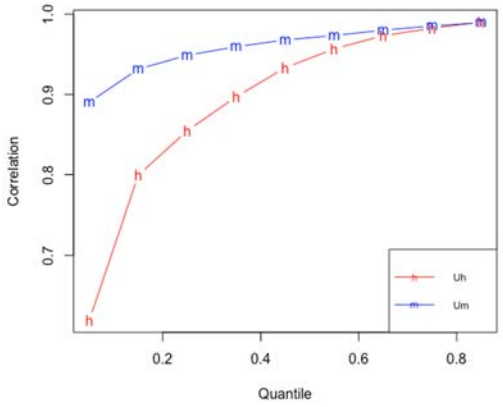


Figure 5. Quantile plot of between topic correlations of the frequency of um *and* uh.

The observations presented so far all suggest that *uh* has greater variability in its distribution in spontaneous conversations compared to *um*. One potential explanation for these observations is the functional or meaning difference between the two fillers. This difference may also relate to the syntactic patterns of different speaker groups.

A potential difference in meaning

In this section, the possibility that the observed difference in frequency variability between the two fillers entails a difference in their meanings is examined through Latent Semantic Analysis (LSA). Essentially, LSA uses the contexts in which a word occurs to represent the 'meaning' of the word of interest. Implementations stemmed from this idea have found wide applications in modern NLP applications.

A word2vec model (Mikolov et al., 2013) was trained with a window size of ±5 words. Metrical Multidimensional Scaling (MDS) was used to project the clusters in word vector space onto the 2D plane for visualization, as plotted in Figure 6. Silent pauses of varying duration were also coded as

different words to explore the correspondence between filler forms and types of delay. Duration thresholds of 400 ms, 600 ms, and 800 ms were arbitrarily set to distinguish between shorter and longer pauses among all the pauses longer than 150 ms.

The first major difference is between silent and filled pauses along the first principle coordinate. In addition, clear difference between the two fillers can also be found along the second principle coordinate, suggesting a distinction of the contexts in which the two fillers potentially occur. More interestingly, the continuum from short to long silent pauses are dispersed along the second principle coordinate, which appears to parallel with the distinction between *uh* and *um* along the same dimension. Therefore it can be argued that there are some contextual similarities between shorter pauses and *uh*, and between longer pauses and *um*.

Discussion and conclusion

In this study I presented evidence for an alternative explanation to the previous observation that the choice of *um* and *uh* as hesitation markers displays consistent age-related variations. Through looking at the absolute frequency variation as a function of age, as well as the effect of topic and interlocutor on individual variation, it can be hypothesized that the so-called change in progress can be at least partially explained by a potential meaning or function difference. This difference may be interpreted as reflecting different aspects in the planning process, which correspond to age-related change in utterance structuring and cognition. This potential difference in meaning or function has found support from LSA. However, substantive proposals for how the meaning and function of the fillers are different is not yet able to be proposed. One obvious possibility is to look at variations in other forms of filler words (such as *be like*) in relation with the two forms of filled pause.

References

Acton, E. K. 2011. On gender differences in the distribution of um and uh. *University of Pennsylvania Working Papers in Linguistics* 17(2): 2.

Cieri, C., D. Miller & K. Walker. 2004. The Fisher Corpus: A resource for the next generations of speech-to-text. In: M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa & R. Silva (eds.), *Proceedings of Language Resources and Evaluation*, Lisbon, Portugal, 69–71.

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

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

Fruehwald, J. 2016. Filled pause choice as a sociolinguistic variable. *University of Pennsylvania Working Papers in Linguistics* 22(2): 41–49.

Goldman-Eisler, F. 1968. *Psycholinguistics: Experiments in spontaneous speech*. London & New York: Academic Press.

Labov, W. 1994. *Principles of Language Change. Volume 1: Internal Factors*. Oxford: Wiley-Blackwell.

Levelt, W. J. 1983. Monitoring and self-repair in speech. *Cognition* 14(1): 41–104. https://doi.org/10.1016/0010-0277 (83)90026-4

Mikolov, T., K. Chen, G. Corrado & J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv*:1301.3781.

O'Connell, D. & S. Kowal. 2005. Uh and um revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research* 34(6): 555–576. https://doi.org/10.1007/s10936-005-9164-3

Rochester, S. R. 1973. The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research* 2(1): 51–81. https://doi.org/10.1007/BF01067111

Tottie, G. 2011. Uh and um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics* 16(2): 173–197. https://doi.org/10.1075/ijcl.16.2.02tot

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

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

The structural signaling effect of silent and filled pauses

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

Abstract

Filled pauses (uh, um) have been shown in a number of studies to have a facilitative effect for listeners, such as helping them better perceive the syntactic structure of ongoing speech. This may be because the extra time afforded by the filled pause gives listeners more time to process the input. Theoretically, then, silent pauses should show a comparable effect. The present study tests this prediction using a grammaticality judgment task following a study by Bailey and Ferreira (2003). Results show that filled and silent pauses have a comparable influence on listeners' grammaticality judgments but further suggest that listeners deem silent pauses as more important and influential markers.

Introduction

The signaling capability of filled pauses in English (i.e. *uh*/*um*) to listeners has been observed in a wide variety of research. Arnold et al. (2003) observed that filled pauses influence listeners' judgments of whether a following noun phrase is a given or new entity in the ongoing discourse. Corley, MacGregor & Donaldson (2006) observed that listeners judged nouns that are immediately preceded by a filled pause as being lower-frequency nouns than those that are not preceded by a filled pause. Brennan and Schober (1999) observed that listeners recover faster when a speech repair is accompanied by a filled pause than when it is not and Bailey and Ferreira (2003) observed that filled pauses influenced listeners' interpretation of the structure of a sentence.

One account of these observations (cf. Bailey & Ferreira, 2003; Corley, MacGregor & Donaldson, 2006; Corley & Stewart, 2008) is what will be referred to here as the *extra time hypothesis*: The additional time afforded to listeners while a filled pause is occurring—as opposed to the contrary case when no filled pause occurs—allows additional linguistic processing to occur, facilitating the listeners' accurate processing of the linguistic input. Evidence consistent with this hypothesis has been obtained in several of the above studies using other sounds in place of filled pauses such as bells, door knocks, or car horns.

However, if the extra time hypothesis is the correct explanation for the various observations, then one would expect that silent pauses—which also afford additional time to the listener for processing—placed where filled pauses might have occurred should show the same effect.

This paper reports on a test of this prediction. The following section reviews previous research on silent and filled pauses and describes the signaling and extra time hypotheses in more detail. The experimental section describes the grammaticality judgment and reaction-time experiment used to evaluate the hypotheses. And the final section discusses the findings and their implications.

Background

Filled pauses

Filled pauses are non-lexical vocalizations uttered by speakers which, by their occurrence, delay the transmission of the linguistic speech signal. In English, where a large proportion of filled pause research has taken place, two basic forms are nearly exclusive: *uh* [əː] and *um* [əːm] (Maclay & Osgood, 1959; Goldman-Eisler, 1961; Mahl, 1987; Shriberg 1994; Clark & Fox Tree, 2002). Many languages have nearly equivalent forms using the same vowel [ə], if available, or a nearby vowel as in French: *euh* [œː], *eum* [œːm] (Vasilescu, Nemoto & Adda-Decker, 2007). Other languages may have somewhat different forms with possibly more than one syllable as in Japanese: *e-* [ɛː], *e-to* [ɛːto] (Maekawa, 2003).

Research on the production and perception of filled pauses in speech has been extensive, covering various academic fields, languages, and investigative paradigms (see Clark & Fox Tree, 2002 for a slightly dated overview). Some of the perceptual research has already been introduced above. Bailey and Ferreira's (2003) work is often cited and underlies the present study. In their work (specifically, their third experiment), they examined the signaling effect of filled pauses relative to clause boundaries in ongoing speech. In a grammaticality judgment task, native English listeners heard sentences as in (1) to (4), in which clause boundaries are marked by brackets.

(1) [Sandra bumped into the uh uh busboy] and [the waiter got angry].
(2) [Sandra bumped into the busboy] and [the uh uh waiter got angry].
(3) [While [the man hunted] the uh uh deer ran into the woods].
(4) [While [the man hunted] the deer uh uh ran into the woods.]

https://doi.org/10.21862/diss-09-006-rose

When the filled pause was at a clause boundary (actually, just after the boundary) as in (2) and (3), listeners were more likely to judge the whole sentence as grammatical than in cases with sentences with a filled pause at a non-boundary as in (1) and (4). Bailey and Ferreira interpreted this as a signaling effect: Listeners interpreted the filled pause as a signal of greater cognitive effort by the speaker, such as that which might be caused by planning a larger speech constituent such as a clause. Thus, they described the filled pauses in (2) and (3) as "good" signals—consistent with the syntactic structure—and those in (1) and (4) as "bad" signals—inconsistent with the structure.

Comparable results were obtained by Bailey and Ferreira when using environmental noises in place of filled pauses. Other researchers have also observed similar effects (e.g. Corley, MacGregor & Donaldson, 2006). Thus, a broader explanation for the observations is the extra time hypothesis. Under this hypothesis, it is the extra time that is afforded by the presence of the filled pause that allows the listener to process the ongoing speech. Hence, when the filled pause is in a position that is consistent with larger constituent boundaries, the listener has time to correctly parse the sentence: the extra time is facilitative. Under this hypothesis, the filled pause is not necessarily a signal, but rather merely an irrelevant filler of time. It is that extra time during which listener processing occurs that causes the observed effects.

Silent pauses

Silent pauses are silent periods which are of unusual duration during a speaker's on-going speech. This excludes short silent periods which may be attributable to breathing or articulatory gestures. Because it is somewhat difficult to determine what counts as "unusual", a common approach in speech studies is to use a somewhat arbitrary threshold. However, even this threshold value has been set widely, as short as 50 ms to as long as 1 sec (De Jong & Bosker, 2013). In recent years, a threshold value around 250–300 ms is common.

Silent pauses, if counted as discrete units of speech, are among the most common tokens in speech corpora (e.g. 3.8 per 100 words in the LOCNEC corpus; Gilquin, 2008). Furthermore, they may play a crucial rule in speech perception. Reich (1980) observed that listeners recalled propositions more accurately when silent pauses were at clause boundary rather than non-boundary positions. These findings might also be explained by the extra time hypothesis: The extra time afforded by the silent pause allows the listener to properly process the structure of the ongoing speech.

Thus, one might predict that filled and silent pauses should show equivalent effects. Yet, another view might be that perhaps the filled pause (or other overt acoustic evidence such as an environmental sound) is not a signal, per se, but simply heightens attention (as discussed in Corley & Stewart, 2008; MacGregor, 2008), perhaps even to the presence of the concurrent delay. This could cause the listener to capitalize on the afforded extra time more effectively.

The present study seeks to investigate these possibilities by replicating Bailey and Ferreira's (2003) study but extending it by comparing the influence of both filled and silent pauses.

Experiment

Materials

The present study used the original stimuli from Bailey and Ferreira (2003) as shown in (1) to (4) with some important changes. In their study, they placed the filled pause between a definite article and its head noun: *the uh uh waiter*; hence, one token away from the relevant clause boundary. This was because their work was building on earlier work on the head noun effect (Ferreira & Henderson 1991) and thus compared these cases to such cases as *the short and pudgy waiter* or *the waiter who was short and pudgy*. However, following their own argumentation, the signaling effect should occur if the filled pause is right at the clause boundary. Thus, in present study, the placement of the filled pauses was as in (5) to (8). Furthermore, only a single filled pause was used rather than the double used in their study. Silent pause stimuli were made by simply replacing the filled pauses with silence.

(5) [Sandra bumped into uh the busboy] and [the waiter got angry].
(6) [Sandra bumped into the busboy] and [uh the waiter got angry].
(7) [While [the man hunted] uh the deer ran into the woods].
(8) [While [the man hunted] the deer uh ran into the woods.]

Their original set of stimuli including coordination (i.e. (5)–(6)) and subordination (i.e. (7)–(8)) constructions was extended to make a total of 90 stimuli. 100 filler stimuli were also used, half of which were ungrammatical in some way.

The stimuli were recorded with a native English speaker's voice (the author's). The filled and silent pauses were acoustically manipulated to ensure that both were a consistent 500 ms long in all stimuli.

Procedure

After completing a consent form, participants were seated in a quiet room with a computer and a pair of comfortable headphones. Stimuli were presented in a randomized manner for each participant using Superlab 4.0 by Cedrus. During each audible stimulus, participants saw a cross "+" fixation symbol. Afterward, they were prompted to judge whether the sentence they heard was grammatical or ungrammatical by pressing a button on the keyboard. Participants were given an explanation of how to judge grammaticality and explicitly told not to judge sentences as ungrammatical merely because of any disfluencies. They also were given sixteen practice items with reinforcing feedback after each. The entire experiment took most participants about 25 minutes.

The recorded data included participants' grammaticality judgments as well as their reaction times, being the time from the offset of the stimulus sentence to the onset of their key press. The data were analyzed using mixed effects modeling with the `nlme` package (v. 3.1-128) in `R` (v. 3.3.2). Fixed effects were signal status (consistent, inconsistent; corresponding to Bailey & Ferreira's, 2003 "good" and "bad" conditions, respectively), gap type (silent, filled) and construction (coordination, subordination). Participants were added to the model as a random effect.

Results

30 university students who were native speakers of North American English participated in the experiment and received US$10 in remuneration.

As illustrated in Figure 1, participants judged the sentence as grammatical more often when the pause was in a consistent signal position (i.e. (6) and (7) above) than when in an inconsistent position (i.e. (5) and (8)) [$t(206) = 8.5$, $p < 0.001$]. Furthermore, an interaction between signal status and gap type [$t(206) = 3.2$, $p < 0.005$] reveals that in the inconsistent signal condition, participants judged the silent pause case less grammatical than the filled pause case. Overall, the mixed effects model explains 29.3% of the variance (marginal R^2).

A significant difference in construction reveals that there is a difference between coordination and subordination stimuli. Hence, Figure 2 shows the breakdown by construction. While the basic pattern of results is unchanged, it is clear that the results are more pronounced with subordination stimuli. Furthermore, post-hoc comparisons of the subordination stimuli show that there is even a difference between the pauses in the consistent condition [$t(29) = 2.1$, $p < 0.05$]: Participants judged the filled pause stimuli as grammatical less often than the silent pause stimuli. This is a reversal of the pattern seen in the inconsistent signal case.

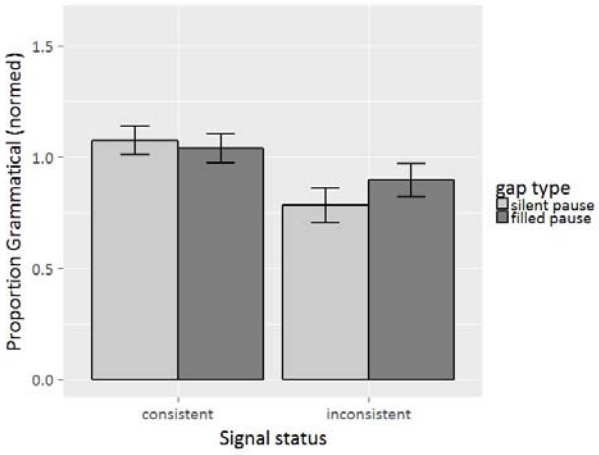


Figure 1. Proportion of stimuli judged grammatical (normed against individuals' judgments). Error bars represent 95% confidence intervals.

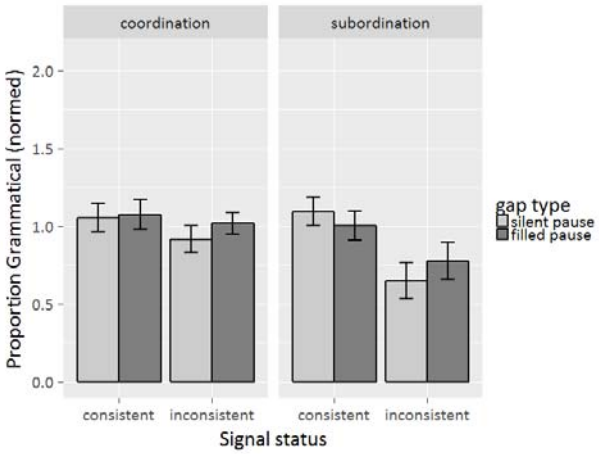


Figure 2. Proportion of stimuli judged grammatical (normed against individuals' judgments) for coordination and subordination stimuli. Error bars represent 95% confidence intervals.

As for the reaction time data (see Figure 3), the trends are in a direction that parallels the grammaticality judgment data: The stimuli in the inconsistent signal condition elicit longer reaction times than those in the consistent condition. Furthermore, the silent pause stimuli yield a slightly longer reaction time than the filled pause stimuli. However, none of these trends reach significance and the marginal R^2 is only 1.2%. The only difference is between the coordination and subordination stimuli with the subordination stimuli showing longer reaction times overall [$t(2127) = 4.7$, $p < 0.001$]. But this difference is not particularly pertinent to the present study.

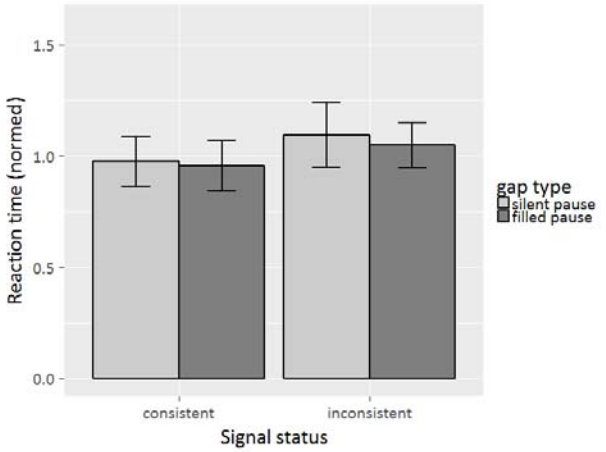


Figure 3. Mean grammatical judgment latency as reaction time (normed against individuals' reaction times). Error bars represent 95% confidence intervals.

Discussion

Results replicate Bailey and Ferreira's (2003) study with respect to filled pause stimuli: Participants judge them less grammatical when in an inconsistent signal position. But the results here further show that a silent pause in an inconsistent signal position is worse, and also that a silent pause in a consistent position is better. The extra time hypothesis is not sufficient to explain this. A partial explanation might come from the idea that the acoustics of the filled pause (like environmental noises) raises the listeners' attention even during the delay time. But this would not quite explain why the silent pause in the consistent condition of subordination stimuli actually yields the highest rate of grammaticality judgment.

In short, the results here suggest that listeners are processing silent pauses and filled pauses each as discrete units of communication and that each signals something a little different. Or, perhaps that they signal the same thing, but at different levels of intensity. But even if so, it would appear that silent pauses—although acoustically nil—are more influential signals—even louder signals—than filled pauses.

Acknowledgments

This work has been partially funded by a Japan Society for the Promotion of Sciences (JSPS) Grant-in-aid (Scientific Research (C), Project #15K02765). I am grateful to Dr. Bonita Miller and Spring Arbor University for their help in carrying out the experiment and also to numerous individuals who have given helpful advice.

References

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

Bailey, K. G. D. & F. Ferreira. 2003. Disfluencies affect the parsing of garden-path sentences. *Journal of Memory and Language* 49(2): 183–200. https://doi.org/10.1016/S0749-596X(03)00027-5

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

Corley, M., L. J. MacGregor & D. Donaldson. 2006. It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition* 105(3): 658–698. https://doi.org/10.1016/j.cognition.2006.10.010

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

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

De Jong, N. H. & H. R. Bosker. 2013. Choosing a threshold for silent pauses to measure second language fluency. In: R. Eklund (ed.), *The 6th Workshop on Disfluency in Spontaneous Speech*, 21–23 August 2013, Stockholm, Sweden, 17–20.

Ferreira, F. & J. M. Henderson. 1991. Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language* 30(6): 725–745. https://doi.org/10.1016/0749-596X(91)90034-H

Gilquin, G. 2008. Hesitation markers among EFL learners: Pragmatic deficiency or difference? In: J. Romero-Trillo (ed.), *Pragmatics and Corpus Linguistics: A Mutualistic Entente*, 119–150. Berlin: Mouton de Gruyter.

Goldman-Eisler, F. 1961. A Comparative Study of Two Hesitation Phenomena. *Language and Speech* 4(1): 18–26. https://doi.org/10.1177/002383096100400102

MacGregor, L. J. 2008. *Disfluencies affect language comprehension: evidence from event-related potentials and recognition memory*. Ph.D. dissertation, The University of Edinburgh.

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

Maekawa, K. 2003. Corpus of Spontaneous Japanese: Its Design and Evaluation. *Proceedings of SSPR*, 13–16 April 2003, Tokyo, Japan, 7–12.

Mahl, G. 1987. *Explorations in nonverbal and vocal behavior*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Reich, S. 1980. Significance of Pauses for Speech Perception. *Journal of Psycholinguistic Research* 9(4): 379–389. https://doi.org/10.1007/BF01067450

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

Vasilescu, I., R. Nemoto & M. Adda-Decker. 2007. Vocalic Hesitations vs Vocalic Systems: A Cross-Language Comparison. In: J. Trouvain (ed.), *Proceedings of 16th International Congress of Phonetic Sciences*, 6–10 August 2007, Saarbrücken, Germany, 1101–1104.

Empathetic hearers perceive repetitions as less disfluent, especially in non-broadcast situations

Iulia Grosman, Anne Catherine Simon and Liesbeth Degand
Institute for Language and Communication, University of Louvain, Louvain-la-Neuve, Belgium

Abstract

This experiment measures the impact of the communicative situation on perceived fluency in French speech. We consider three dimensions of fluency: grammatical, discursive and socio-interpersonal. We first hypothesise that grammatical fluency is less influenced by contextual constraints than the other two dimensions. Furthermore, taking into account the Interpersonal Reactivity Index of each participant, we hypothesise that hearers with higher interpersonal capacities will be more tolerant in their fluency evaluation, because of their ability to project into the speaker's mind. The strength of the design rests on the proposal to test natural stimuli and integrate social and individual variables in a perception experiment.

Introduction

Evaluation of fluency

This study evaluates three types of fluency: the grammatical dimension of fluency, the discourse-level one and the socio-interpersonal one. This tri-dimensional model of fluency, proposed in Grosman (2018), is an expansion of the socio-cognitive framework on saliency proposed by Schmid and Günther (2016).

In this model, a stretch of speech is evaluated as being disfluent when the (grammatical, discursive, socio-interpersonal) discourse expectations of the hearer are over-confirmed or over-deceived. This evaluation depends on the degree of convergence of the discourse with the hearers' expectations.

The first hypothesis is that grammatical fluency is less dependent on contextual constraints than discursive and socio-interpersonal fluency. The second hypothesis is that the higher the ability to project into a speaker's mind, the more lenient hearers are in their evaluation.

We modulate the production of repetitions to evaluate the perceived dimension of fluency. We focus solely on identical repetitions structure that interrupts the syntactic, grammatical speech flow, and might cause socio-cognitive dis/fluency. Four conditions for *repeated segments* were considered: contiguous, with a filled pause, with a silent pause and no repetition.

Method and design

For each participant, the collected data comprises the results of a pretest (Interpersonal Reactivity Index questionnaire) and the degree of agreement with 6 assertions related to 40 speech stimuli. Furthermore, the experiment tests the impact of the broadcasting orientation on fluency evaluation and on the different dimensions of fluency.

Participants

The 202 native francophone participants were allowed to use hi-fi speakers or headphones (43% and 57% respectively). Their age ranged from 19 to 72 ($\mu = 30.62$, $\sigma = 11.62$) and 69% of them were women. Almost half of the participants were bilingual (47%) and only a quarter declared themselves as monolingual or trilingual (26% and 23% respectively).

Interpersonal Reactivity Index (F-IRI)

The participants completed the F-IRI test (Davis, 1983; Gilet et al., 2013). Through a 28-item self-report, the index evaluates aspects of empathy: the capacity to put oneself in someone else's shoes, to imagine his/her thoughts and feelings.

The test is structured around four subscales: (1) *Perspective Taking*: tendency to adopt the point of view of other people in everyday life; (2) *Fantasy*: tendency to transpose oneself into the feelings and actions of fictitious characters in books, movies, and plays; (3) *Empathic Concern*: tendency to experience feelings of warmth, compassion and concern for other people; (4) *Personal Distress*: assessment of typical emotional reaction of personal unease and discomfort in reaction to the emotions of others.

The mean score of F-IRI is 95.6 ($\sigma = 12.3$, $max = 123$, $min = 60$). The subscale Empathic Concern reaches the highest mean score ($\mu = 26$, $\sigma = 4.7$). It is followed by the subscales fantasy ($\mu = 25.9$, $\sigma = 5.2$), Perspective Taking ($\mu = 24.5$, $\sigma = 4.6$) and Personal Distress ($\mu = 19.24$, $\sigma = 5.4$). We categorised the participants' scores into 5 quartiles for each subscale (e.g. lowest, low, average, high and highest scores).

Audio stimuli selection and processing

The experiment tests the impact of repetitions on perceived fluency (see examples below). We consider four conditions for repeated segments: contiguous,

https://doi.org/10.21862/diss-09-007-gros-etal

with a filled pause, with a silent pause and no repetition. The 40 natural stimuli were extracted from C-HUMOUR (Grosman, 2016), LOCAS-F (Degand, Martin & Simon, 2014), C-Phonogenre (Goldman, Prsir & Auchlin, 2014), Rhapsodie (Kahane & Peitrandrea, 2019) and DRIVE (Christodoulides, 2016).

We selected semi-automatically two sets of original speech samples: 20 original items including a contiguous identical repetition (REP-C) and 20 original items including an identical repetition with a filled pause in the editing phase (REP-FP). We transformed both sets respectively into 20 modified repetition-free items, by removing the first segment (NO-REP), and 20 modified repetition items with a silent pause (REP-UP). A participant listening to a speech sample in the REP-C or REP-FP condition would never hear its counterpart in the NO-REP or REP-UP condition, and vice versa.

Examples:

- REP-C to No-REP: ...la ville est assez haute et {(la) la} Loire est en bas...
  ...the city is quite high and {(the) the} Loire is below...) (1521-2-2NM-Rhap-D0003)
- REP-FP to REP-UP: ...ces nobles vont en général chanter entre guillemets {leur (euh)|(//) leur} nostalgie...
  ...the nobles will generally sing {their (uh)|(//) their} nostalgia... (386-1-2NM-conv-f-2)

Based on the most frequent occurrences in the distribution of repetitions in the corpus data (Grosman et al., 2017), we selected monosyllabic, single repetitions of grammatical determiners. We excluded repetitions occurring at the very beginning or very end of a sample, avoiding primacy or recency effect during evaluation. Speech sample duration varies from 7 to 19 sec. ($\mu = 11.3$, $\sigma = 3.1$).

Impact of broadcasting orientation

Each stimulus was contextualised by an image indicating the communicative situation: broadcast oriented or non-broadcast-oriented speech (see Figure 1). The speech samples were either presented in their original production setting or in a pretended one. Each participant evaluated 40 stimuli, 10 stimuli by fluency condition, with half of them presented in their original context of production and the other half in the pretended modified context.

Each stimulus was paired with a short nominal sentence of contextualisation (e.g. Writing a book on Jimmy Hendricks, Choosing a job, Spider reproduction). The experiment fosters hearers to focus on the speech form rather than on its content.

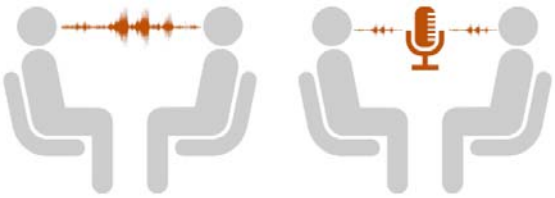


Figure 1. The contextualisation images of a broadcast and non-broadcast-oriented speech.

Tridimensional evaluation of fluency

For each of the 40 stimuli listened to, participants expressed their degree of agreement with 6 assertions (40 stimuli × 6 assertions = 240 answers per participant). Each pair of assertions corresponded to an evaluated dimension of fluency described above:

- *Grammatical:*
  A1. The sentence is well-formed.
  A2. The sentence includes hesitations.
- *Discourse-level:*
  A3. The speech is fluid.
  A4. The speech is nice to listen to.
- *Socio-interpersonal:*
  A5. The speech in this context appeals to me.
  A6. The speech used in this context is improper.

For the evaluation of the results, the scale of A2 and A6 assertions were inverted as they were negatively formulated. The distance between each ordinal point was considered equidistant as there was a significant information value added compared to a model with symmetric thresholds (around the central one, here "average") ($LRtest = 535.25$, $p < 0.001$).

Hypothesis

In this framework, disfluency corresponds to an over-saliency by converging or diverging expectations of the hearer regarding context. Depending on the context considered, distinct (dis)fluency evaluations emerge from different types of expectations. Regarding fluency evaluation, the main hypothesis is that the evaluation of grammatical fluency is less conditioned by contextual constraints than the evaluation of discursive fluency. Some specific hypotheses (A–C) were also tested.

A. *Hypothesis on Internal Structure of the Repetition:* The internal structure of a repetition influences its evaluation, independently of the discursive context. More precisely, non-contiguous repetitions with an editing term (*the uh the*) are evaluated less leniently than contiguous repetition (*the the*). This hypothesis supports the following continuum: (fluent) > no repetition > repetition with filled pause > repetition with unfilled pause > contiguous repetition > (disfluent).

The orientation of this hypothesis is motivated in twofold way. Results observed in production reveal

the relative importance of repetition involving a silent pause, even in prepared discourse (Grosman, 2018: 252–265). Results observed in perception reveal the ease of processing and the saliency accordingly for each structure (Grosman, 2015).

B. *Hypothesis on Broadcast Speech:* The stimuli presented in a formal broadcast-oriented speech context are rated more severely than those in non-broadcast-oriented speech. Additionally, a same repetition in one context of production does not necessarily induce the same evaluation for every hearer. We do not formulate any hypothesis linking the type of repetition and the context of production (i.e. one structure is not more accepted than another in a certain context), as this latter feature was non-relevant in our corpus data (Grosman, 2018: 252–265).

C. *Hypothesis of F-IRI:* The higher the interpersonal capacities of the hearers are, the more lenient they may be in their evaluation, due to the ability to project into the speaker's mind. Furthermore, we think the F-IRI is more relevant when evaluating discursive or socio-interpersonal fluency than grammatical fluency.

Finally, the perception and evaluation of disfluency depend on one's ability to anticipate content. The better the subject's projection capacity is, the more s/he can process (dis)fluency efficiently. Hence, subjects getting high or very high scores in the dimension of Empathic Concern, Perspective Taking and Imagination would rate fluency less harshly.

Main Results

Importance of stimuli, participant and thresholds

The results were processed with an ordinal mixed model with 2 random effects recognizing the fact that (1) the same audio stimulus was evaluated according to 6 fluency measures and (2) the same participant evaluated 40 stimuli. The compared results between the ordinal model and the ordinal mixed model showed that 22.44% of the variance is explained by these random effects ($Pseudo\text{-}R^2 = 0.224$, $p < 0.001$).

Internal structure of the repetition

The cumulative mixed effect model reveals that the overall degree of perceived fluency tends to lower when a repetition appears, no matter the internal structure of the repetition (Table 1). A stimulus with a repetition has 88% probability to be evaluated as more disfluent that the repetition-free ones.

Results show that the force of the devaluation is stronger when evaluating the socio-interpersonal dimension of fluency ($OR = 0.97$, $p < 0.001$) than for the discursive or grammatical dimensions (respectively $OR = 0.86$ and $OR = 0.88$ at $p < 0.001$). Hence, a speech sample including a repetition was not considered as including more hesitation or being inappropriate but rather as less pleasant to listen to and less fluid.

Table 1. Fixed effect results on the fluency score depending on the type of repetition (NO-REP as intercept).

(Intercept)	*Estimates*	*OR*	*ES*	*z*	*P*
REP-C	−0.126	0.889	0.024	−5.29	< 0.001
REP-FP	−0.166	0.848	0.218	−0.76	0.446
REP-UP	−0.138	0.872	0.218	−0.63	0.526

Impact of broadcast speech situation

Overall, there is a significant difference in the appreciation of perceived fluency depending on the situation in which the speech is produced (*est.* = 0.33, $OR = 1.03$, $p < 0.05$). The degree of perceived fluency increases in broadcast situation. The participants rate less severely non-broadcast situations than broadcast ones. Nevertheless, no interaction was found with any dimension of fluency, the supposed situation impacting every dimension equally.

Impact of interpersonal reactivity

Results show that F-IRI scores do not have an impact on overall fluency perception (*est.* = 0.006, $p < 0.15$), but that it does influence each dimension of fluency separately. In fact, the higher the F-IRI scores of participants are, the higher they evaluate the samples on the discursive and socio-interpersonal dimensions of fluency (respectively *est.* = 0.39, $p < 0.03$ and *est.* = 0.27, $p < 0.005$), but not on the grammatical one (*est.* = −0.04, $p < 0.7$).

In addition, each subscale of the F-IRI test does have an impact on overall fluency evaluation and on its different dimensions. Overall, the more participants show empathic concern, the more they rate items as fluent (*est.* = 0.33, $p < 0.03$). Similarly, the more one shows personal distress, the more one judges the items as fluent (*est.* = 0.29, $p < 0.02$).

The evaluation of *grammatical fluency* is mainly influenced by the participant's imagination (*Fantasy*) (*est.* = −0.34, $p < 0.03$). The more participants have imagination, the more fluent they judge the grammatical aspect of a speech sample.

The evaluation of *discourse-level fluency* is mainly influenced by the empathetic concern of a participant (*est.* = 0.41, $p < 0.01$). Confirming our hypothesis, the more participants show empathetic concerns, the more they rate a speech sample in its context as fluent.

The evaluation of *socio-interpersonal* fluency is mainly impacted by the participants' personal distress

($est. = 0.38$, $p < 0.02$). The more one has personal distress capacity, the more one evaluates a speech sample as fluid and enjoyable to listen to.

Discussion and conclusion

This study unveils three main results:

Regarding the first hypothesis, the internal structure of repetitions does not impact the overall perceived fluency. The presence of a repetition favours by itself the perception of a speech sample as less fluent, especially when evaluating the socio-interpersonal dimension (more than the grammatical or discourse-level ones). A speech sample is not perceived as including more hesitation or being inappropriate, but rather as less pleasant to listen to and less fluid.

Regarding the second hypothesis, participants overall rate less severely repetitions in non-broadcast situations than in broadcast ones, supporting the idea that the type of speech evaluated by the participants matters greatly.

Regarding the third hypothesis, the F-IRI test shows that the higher the scores of participants are, the more lenient they are to rate items as fluent on the discursive and socio-interpersonal dimensions. The overall F-IRI does not have an impact on the grammatical evaluation of fluency, but each subscale of the F-IRI brings refinements to the analysis. The grammatical fluency is particularly impacted by the participants' imagination, while the discursive fluency evaluation is mainly influenced by their empathetic concern, while the socio-interpersonal fluency is impacted by their level of personal distress.

This paper broadly questions (dis)fluency perception and evaluation: the type of discourse that is under evaluation, the empathetic participants profile and their personal discourse expectations, as well as the type of fluency evaluated. It reveals that certain types of evaluation (e.g. interpersonal fluency) require to consider individual profiles more than others (e.g. grammatical fluency).

In this sense, it addresses several challenges: (1) the benefits of coupling production and perception data for an integrated study of repetitions; (2) the advantage of using natural aligned speech data, selected on the basis of corpus analysis, for perceptive experiments; (3) the importance of integrating individual traits into fluency evaluation; (4) the impact of the speech situation on fluency evaluation.

To address these challenges, we proposed an integrated model of fluency according to which the evaluation of fluency depends on the degree of convergence/divergence between the discourse and the hearer expectations.

Acknowledgements

This research benefits from the support of the ARC-project "A Multi-Modal Approach to Fluency and Disfluency Markers" granted by the Fédération Wallonie-Bruxelles (grant nr. 12/17-044). We thank S. Roekhaut, L. Rousier-Vercruyssen and G. Christodoulides for technical support.

References

Christodoulides, G. 2016. *Effects of cognitive load on speech production and perception*. Ph.D. dissertation, University of Louvain, Belgium.

Degand, L., M. Martin & A. C. Simon. 2014. LOCAS-F : un corpus oral multigenres annoté. In: *Congrès Mondial de Linguistique Française*, Berlin, Germany, 2613–2625. https://doi.org/10.1051/shsconf/20140801211

Davis, M. H. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology* 44(1): 113–126. https://doi.org/10.1037/0022-3514.44.1.113

Gilet, A.-L, M. Mella, J. Studer, D. Grühn & G. Labouvie-Vief. 2013. Assessing dispositional empathy in adults: A French validation of the Interpersonal Reactivity Index (IRI). *Canadian Journal of Behavioural Science* 45(1): 42–48. https://doi.org/10.1037/a0030425

Goldman J.-P., T. Prsir & A. Auchlin. 2014. C-PhonoGenre: a 7-hour Corpus of 7 Speaking Styles in French. In: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.): *Proceedings of Language Resources and Evaluation*, 26–31 May, Reykjavik, Iceland.

Grosman, I. 2015. Complexity Cues or Attention Triggers? Repetitions and Editing Terms for Native Speakers of French. In: *Proceedings of Disfluency in Spontaneous Speech 2015,* Edinburgh, UK.

Grosman, I., C. Christodoulides, L. Degand & A.-C. Simon. 2017. Prosodic Variation of Identical Repetitions as a Function of their Properties and Editing Terms. In: *International Conference on Fluency and Disfluency Across Languages and Language Varieties*. Louvain-la-Neuve, Belgium.

Grosman, I. 2018. *Évaluation contextuelle de la (dis)fluence en production et perception : pratiques communicatives et formes prosodico-syntaxiques en français* [Contextual evaluation of (dis)fluency in production and perception: Communicative practices and prosodico-syntactical forms in French]. Ph.D. dissertation, University of Louvain, Belgium.

Kahane S. & P. Pietrandrea. 2019. The syntactic annotation of the Rhapsodie corpus: an overview. In: A. Lacheret, S. Kahane & P. Pietrandrea (eds.), *Rhapsodie: a prosodic syntactic treebank of spoken French*. Amsterdam & Philadelphia: John Benjamins Publishing Company, 35–47. https://doi.org/10.1075/scl.89.04kah

Schmid H.-J. & F. Günther. 2016. Toward a Unified Socio-Cognitive Framework for Salience in Language. *Frontiers in Psychology* 7, 1110. https://doi.org/10.3389/fpsyg.2016.01110

Pausing strategies with regard to speech style

Dorottya Gyarmathy and Viktória Horváth
Department of Phonetics, Research Institute for Linguistics,
Hungarian Academy of Sciences, Budapest, Hungary

Abstract

Speech is occasionally interrupted by silent and filled pauses of various length. Pauses have many different functions in spontaneous speech (e.g. breathing, marking syntactic boundaries as well as speech planning difficulties, time for self-repair). The aim of the study was the analysis of the interrelation between the temporal pattern and the syntactical position of silent pauses (SP) on one hand. On the other hand, filled pauses (FP) were also analyzed according to their phonetic realization, as well as the combination of SPs and FPs. The effect of speech style on pausing strategies was also analyzed. A narrative recording and a conversational recording from 10 speakers (ages between 20 and 35 years, 5 male, 5 female) were selected from Hungarian Spontaneous Speech Database for the study. The material was manually annotated, silent pauses were categorized, then the duration of pauses were extracted. Results showed that the position of silent and filled pauses affects their duration. The speech style did not influenced the frequency of pauses. However, silent and filled pauses were longer in narratives than in conversations. Results suggest that pausing strategies are similar in general; however, the timing patterns of pauses may depend on various factors, e.g. speech style.

Introduction

Pauses serve various functions in speech, like breathing, grammatical function, providing time for speech planning processes and for perception as well (Levelt, 1989, Gósy, 2000). The realization of pauses depends on various factors, e.g. the speaker's age, the length and the complexity of the utterance or the speech style (Duez, 1982, Krivokapic, 2007). Researches revealed connection between the speech situation and the pauses. The more complex a speech task was—the greater cognitive effort it required—the longer and more frequent the pauses became (Goldman-Eisler, 1968; Kowal et al., 1975). Silent pauses were longer and more frequent in political speech, the longest pauses having a stylistic function. Filled pauses were not characteristic for this type of speech, whereas they were decidedly frequent in interview situations (Duez, 1982). Connection was found between the position and the duration of pauses, for example in the case of 'to+infinitive' grammatical structures. There were significantly longer pauses before 'to' than after it during reading aloud; whereas the opposite was found in spontaneous speech, probably due to speech planning characteristics (Bada & Genç, 2008).

The effect of speech style on pausing were analyzed in Hungarian as well. Researches revealed differences in pausing strategies between spontaneous speech and reading aloud (Olaszy, 2005; Váradi, 2010). The ratio of pauses was less in conversations than in narratives, in addition the duration of pauses was shorter in conversations than in narratives (Markó, 2005). Re-telling a story was the most difficult speech task for young adults; therefore, the speakers produced pauses the most frequently in this task. The pauses realized with longer duration in re-telling a story than in conversations (Bóna, 2013).

Silent and filled pauses have several additional functions in conversations (Sacks et al. 1974). 'Pause' is defined as a signal break within a speech turn (we analyzed this type of silent pauses in the present study). Furthermore, pauses occur in conversations for thinking or for dramatic effect, the speaker can use them to highlight new information, and they can also be used to organize the discourse (Esposito et al., 2007).

The aim of the study is to analyze the occurrence and duration of silent pauses according to their position in conversations and narratives, on one hand. On the other hand, the realization of filled pauses and their combination with silent pauses were also analyzed. Our hypotheses were that (i) silent and filled pauses realize with different patterns according to speech style; (ii) the duration of silent and filled pauses is determined by their position.

Method and material

10 conversations and 10 narratives were selected for the study from a Hungarian Database called BEA (Neuberger et al., 2014). Three speakers participated in each conversation; the interviewer (Int) and one speaker (henceforth: the second speaker S2) were constant. S1 speaker was the third participant in each conversation. The S1 speaker was asked to tell his or her opinion on a given topic by the interviewer in the narrative sessions. The S1 participants were between 20 and 35 years old, half of them were male and half of them were female. Both the Int and S2 were 28 years old.

https://doi.org/10.21862/diss-09-008-gyar-horv

The total material was 175 minutes long (conversations altogether 131 minutes, *mean* = 13 min., range: 6.9–23.3 min.; while narratives altogether 44.5 minutes, *mean* = 4.45 min, range: 1.7–10.2 min). The annotation was carried out in Praat (Boersma & Weenink, 2018). The speech intervals, the silent and filled pauses were annotated, labelling the phonetic form of filled pauses as well. The duration of silent and filled pauses were automatically extracted.

1853 silent pauses occurred in the total material, 1185 in the conversations, while 668 in the narratives. Silent pauses were categorized based on the system developed by Gyarmathy (2018). The first distinction was whether the pause was related to disfluency (in these cases, the time span between the interruption of articulation and the beginning of correction was taken into account, as part of the editing phase – E), or it had a syntactical function (S) (Figure 1). Pauses with editing function (E) were further categorized based on whether the disfluency phenomena were due to the speaker's uncertainty or errors. Silent pauses with a syntactical function (S) were distinguished based on their position. Utterance onset pauses (S_Uo) occur when a speaker claims the turn; here the pause may only be preceded by a filler word or a discourse marker. Silent pauses at phrase boundaries (S_PhrB) are found between clauses of virtual sentences, often before or after a conjunction. Within phrase pauses (S_PhrW) are found within a grammatical unit ('phrase'). End of phrase pauses (S_PhrE) are silent pauses at the end of a virtual sentence, after which the speaker starts another virtual sentence that often represents a new thought unit. The frequency and the duration of silent pauses was also analyzed with regard to these categories. 555 filled pause occurred in the total material. The phonetic form, the frequency, the position and the duration of filled pauses were also analyzed, as well as the occurrences of FPs combined with silent pauses. Statistical analysis was conducted using SPSS 20.0 (GLM, GLMM).

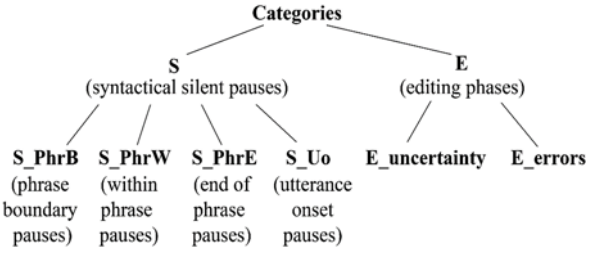


Figure 1. The categories of intra-speaker silent pauses.

Results

Results showed that the silent pause was the most frequent type, irrespectively of the speech style (5.8 items per 100 syllables in conversations on average, while 5.6 in narratives, see Figure 2).

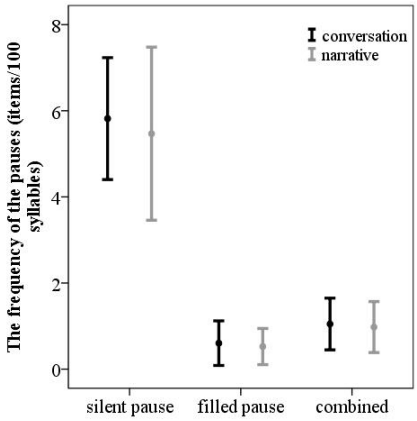


Figure 2. The frequency of pauses according to speech style.

The occurrence of combined pauses was about 1 item per 100 syllables on average, while the mean frequency of filled pauses was 0.5 item, irrespectively of the speech style.

The duration of silent pauses (Figure 3) were significantly longer in the narratives (523 ms on average) than in the conversations (466 ms on average) [GLMM: $F(1, 1851) = 10.057$ $p = 0.002$ pairwise: $t = 3.171$].

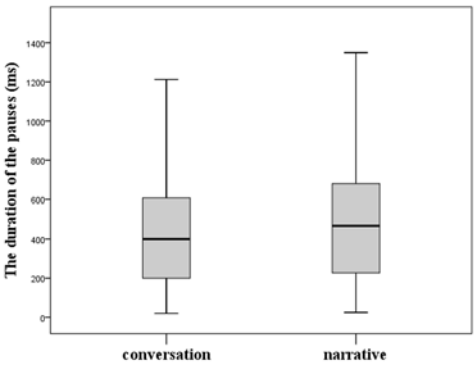


Figure 3. The duration of the silent pauses according to the speech style.

The occurrence and the duration of SPs were also analyzed according to the position. The most of the silent pauses were syntactical pauses (S) with the occurrences of 5 items per 100 syllables in the conversations and in the narratives as well. Pauses in editing phases (E) of disfluencies were considerably less frequent: 0.7 items per 100 syllables in conversations, while 0.5 items in narratives. The ratio of the subcategories of syntactical silent pauses were similar in both of the speech styles. Pauses occurred at phrase boundaries most frequently (S_PhrB): 2.6 items per 100 syllables in conversations and in narratives as well. 1.3 pauses

occurred per 100 syllables within phrases (S_PhrW)—the frequencies were the same in both of the speech styles. Pauses at the utterance onset (S_Uo) occurred the least frequently (only 0.05 items per 100 syllables in conversations and 0.04 in narratives), irrespectively of speech style. Silent pauses in editing phases occurred somewhat more frequently in the conversations: 0.58 items connected with uncertainty and 0.16 items connected with errors per 100 syllables (narratives: 0.46 items of E_unc and 0.09 items of E_errors).

We also analyzed the duration of pauses according to the subcategories (Figure 4.). The longest silent pauses were the S_PhrE and the S_Uo types, irrespectively of speech style. The S_Uo pauses were longer in conversations (697 ms, *SD* = 999 ms) than in narratives (500 ms, SD = 647 ms). In contrast, pauses at the end of the phrases (S_PhrE) were longer in narratives (754 ms, *SD* = 479 ms) than in conversations (677 ms, *SD* = 430 ms). The S_PhrB pauses were longer than S_PhrW pauses in conversations and in narratives as well (conversations: S_PhrB: 449 ms, *SD* = 329 ms; S_PhrW: 372 ms, *SD* = 259 ms; narratives: S_PhrB: 529 ms, *SD* = 366 ms; S_PhrW: 367 ms, *SD* = 275 ms). The pauses of editing phases connected to uncertainty phenomena (S_unc) realized with longer durations than pauses of editing phases connected to errors, irrespectively of speech style (S_unc in conversations: 388 ms, *SD* = 324 ms; in narratives: 418 ms, *SD* = 340 ms; E_error in conversations: 306 ms, *SD* = 314 ms; in narratives: 159 ms, *SD* = 170 ms). The statistical analysis revealed that the subcategories of silent pauses determine their duration in conversations [GLMM: $F(5, 1834) = 24.794$, $p < 0.001$] and in narratives [GLMM: $F(5, 1834) = 22.496$, $p < 0.001$] as well (Table 1 contains the results of pairwise contrasts).

The filled pauses realized with various phonetic forms. They consisted either of one speech sound or of two or three speech sounds. The single speech

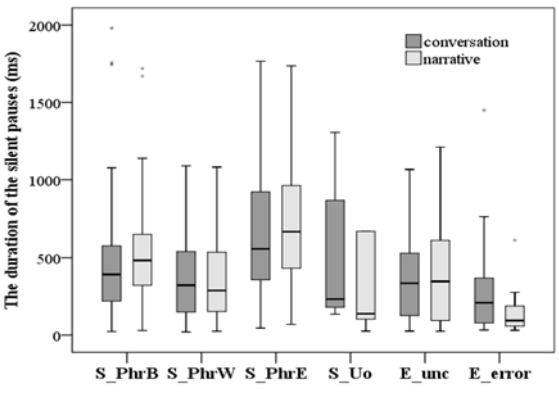


Figure 4. The duration of the silent pauses according to subcategories and speech style.

Table 1. The results of the pairwise contrast according to speech style.

speech style	subcategories of silent pauses	*t*-value	significance contrast
conversation	**S_PhrB – S_PhrW**	3.354	0.001
	S_PhrB – S_PhrE	8.108	0.000
	S_PhrB – E_unc	2.734	0.006
	S_PhrB – E_error	2.455	0.014
	S_PhrW – S_PhrE	9.889	0.000
	S_PhrE – E_unc	8.126	0.000
	S_PhrE – E_error	5.750	0.000
narrative	**S_PhrB – S_PhrW**	4.596	0.000
	S_PhrB – S_PhrE	6.375	0.000
	S_PhrB – E_unc	2.767	0.006
	S_PhrB – E_error	3.021	0.003
	S_PhrW – S_PhrE	9.391	0.000
	S_PhrE – S_Uo	3.039	0.002
	S_PhrE – E_unc	6.658	0.000
	S_PhrE – E_error	5.018	0.000

sound was a neutral vowel or a bilabial nasal-like consonant. Filled pauses consisting of more speech sounds were combinations of schwa and nasal with each other or with laryngeal consonant. The schwa was the most frequent form in eight speakers' speech (72% on average, 55–100%) in conversations, while the ratio was 72% on average in narratives (51–100%). We analyzed the combined occurrences of silent and filled pauses. Filled pauses occurred between two words (without any silent pauses) in a similar ratio than preceding a silent pause, irrespective of the speech style (Figure 5). The less frequent case was when filled pauses occurred between two silent pauses.

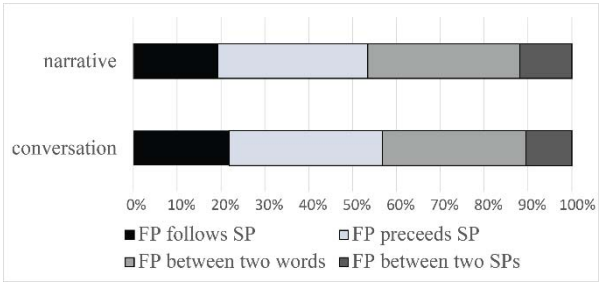


Figure 5. The combined occurrences of silent and filled pauses.

The duration of the most frequent filled pause (schwa) was analyzed with regard to its position and the speech style. The position of schwa had statistically significant effect on its duration [GLM: $F(1, 277) = 19.678$ $p = 0.001$ $\eta^2 = 0.664$]. Filled pauses were significantly longer between two silent pauses (*mean* = 434 ms, *SD* = 211 ms) than the filled pauses adhered to word(s) (*mean* = 296 ms, *SD* = 141 ms). The speech style also influenced the

duration of schwa-like FPs [GLM: $F(1, 277) = 5.322$ $p = 0.025$ $\eta^2 = 0.099$]. The duration of filled pauses were significantly longer (*mean* = 343 ms, *SD* = 154 ms) in the narratives than in the conversations (*mean* = 295 ms, *SD* = 155 ms).

Conclusions

Silent and filled pauses were analyzed in the study, according to the speech style. Our first hypothesis were partly confirmed: the frequency of SPs (and its subcategories as well) and FPs was similar in conversations and narratives of the certain speakers. However, SPs and FPs were significantly longer in narratives than in conversations, similar to earlier studies (cf. Markó, 2005). The task of speaking is easier in conversation than in narratives, due to the fact that partners help each other in managing the conversations on one hand. On the other hand, during the partner's speech, the following speaker has time for speech planning processes. Silent pauses appeared more often in grammatically functional positions in total (S_PhrB, S_PhrE, S_Uo) than within a phrase, irrespectively of speech style. Speakers usually do not create a break within a utterance; this indicates they not only plan the content and form of the utterance, but also the pauses (cf. Zellner, 1994). Within-phrase pauses can be a sign of a major speech planning problem. Data confirmed our hypothesis: the position of SPs determined their duration: pauses at boundaries were longer than pauses within a phrase. S_PhrBs were shorter in conversations than in narratives—their shorter duration may mark the speaker's intention of keeping the floor. The position of FPs influenced their realization. FPs occurred the least frequently between two silent pauses; however, with the longest durations in this position – they presumably indicate problems in planning processes.

Acknowledgements

The research was supported by the Hungarian National Research, Development and Innovation Office of Hungary [projects No. K-128810] and the Bolyai János Research Scholarship.

References

Bada, E. & B. Genç. 2008. Pausing preceding and following to in to-infinitives: A study with implications to reading and speaking skills in ELT. *Journal of Pragmatics* 40(11): 1939–1949. https://doi.org/10.1016/j.pragma.2008.03.010

Boersma, P. & D. Weenink. 2018. *Praat: Doing phonetics by computer* (version 6.0.19). http://www.praat.org/ (accessed 24 May 2019).

Bóna J. 2013. A beszédszünetek fonetikai sajátosságai a beszédtípus függvényében [Phonetic features of pauses depending on the type of speech]. *Beszédkutatás* 21: 60–75.

Duez, D. 1982. Silent and non-silent pauses in three speech styles. *Language and Speech* 25(1): 11–25. https://doi.org/10.1177/002383098202500102

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

Goldman-Eisler, F. 1968. Psycholinguistics: *Experiments in spontaneous speech*. London: Academic Press.

Gósy, M. 2000. A beszédszünetek kettős funkciója [The double function of pauses in speech]. *Beszédkutatás* 2000: 1–14.

Gyarmathy, D. 2018. The functions of silent pauses in spontaneous Hungarian speech. *The Phonetician* 115: 53–71.

Kowal, S., D. C. O'Connell & E. J. Sabin. 1975. Development of temporal patterning and vocal hesitations. *Journal of Psycholinguistic Research* 4(3): 195–207. https://doi.org/10.1007/BF01066926

Krivokapic, J. 2007. Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics* 35(2): 162–179. https://doi.org/10.1016/j.wocn.2006.04.001

Levelt, W. J. M. 1989. *Speaking: From intention to articulation*. A Bradford Book. Cambridge (Massachusetts)–London (England): The MIT Press.

Markó, A. 2005. *A spontán beszéd néhány szupraszegmentális jellegzetessége* [Some suprasegmental features of spontaneous speech]. Ph.D. dissertation, ELTE, Budapest.

Neuberger, T., D. Gyarmathy, T. E. Gráczi, V. Horváth, M. Gósy & A. Beke. 2014. Development of a large spontaneous speech database of agglutinative Hungarian language. In: P. Sojka, A. Horák, I. Kopeček & K. Pala (eds.), *Proceedings of TSD 2014 Text, Speech and Dialogue, the 17th International Conference*, 424–431. https://doi.org/10.1007/978-3-319-10816-2_51

Olaszy, G. 2005. Prozódiai szerkezetek jellemzése a hírfelolvasásban, a mesemondásban, a novella és a reklámok felolvasásában [Prosodic features in news items, storytelling, short stories and commercials]. *Beszédkutatás* 2005: 21–50.

Sacks, H., E. A. Schegloff & G. Jefferson. 1974. A simplest systematics for the organization of turn taking for conversation. *Language* 50(4/1): 696–735. https://doi.org/10.1353/lan.1974.0010

Váradi, V. 2010. A felolvasás és a spontán beszéd temporális sajátosságainak összehasonlítása [Comparing the temporal features of reading aloud and spontaneous speech]. *Beszédkutatás* 2010: 100–109.

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

The effects of read-aloud assistance on second language oral fluency in text summary speech

Shungo Suzuki and Judit Kormos
Department of Linguistics and English Language, Lancaster University, UK

Abstract

Focusing on text summary speaking tasks, the present study investigated the effects of the activation of phonological representations during text comprehension (operationalized by read-aloud assistance) on the subsequent retelling speech. A total of 24 Japanese learners of English completed text summary speaking tasks under two conditions: (a) reading without read-aloud assistance and (b) reading with read-aloud assistance. Their speech data were analyzed by lexical overlap indices (i.e. the ratio of characteristic single-words and multiword sequences) and by fluency measures capturing three major dimensions of fluency—speed, breakdown, and repair fluency. The results showed that read-aloud assistance directly facilitated lexical overlaps with source texts and indirectly improved speed and repair fluency. Furthermore, read-aloud assistance was found to affect the interrelationship between lexical overlaps and utterance fluency. The findings suggested that read-aloud assistance might help second language learners to store multiword sequences as a single unit (i.e. chunking) during text comprehension.

Background

Previous studies suggest that the activation of speech content and linguistic forms prior to speech should contribute to efficient speech processing, leading to fluent speech (Skehan, 2014). From the perspective of L2 speech production, the processes of specifying speech content and linguistic forms (i.e. *conceptualization* and *formulation*) mainly rely on controlled processing (Kormos, 2006). Thus, the activation of underlying speech processing can be assisted at the following different levels: content specification, lemma selection, syntactic structure, and phonological representations.

One of the effective fluency enhancement strategies is the use of multiword sequences (MWS) including formulaic sequences and n-grams (Stengers et al., 2011). L2 speech production model assumes MWS are stored as a whole and are retrieved as a single unit. This direct single-step retrieval can save attentional resources for other speech processing such as conceptualization or syntactic and phonological encoding (Skehan, 2014).

Considering the advantages of the activation of linguistic forms as well as the use of MWS, L2 speakers tend to be fluent in text summary tasks where they retell what they read. In addition, read-aloud assistance (RAA), in which one reads a text while simultaneously listening to its oral recording, may further enhance their utterance fluency (UF) due to its dual-modal input. For instance, intonation can help L2 learners to segment texts into larger units of grammatical and/or semantic information (cf. Košak-Babuder et al., 2019). Therefore, RAA can be hypothesized to lead to the facilitated text comprehension as well as the activation of MWS (i.e. *chunking)*. However, it is still unclear the extent to which such an enhanced reading by dual-modal input can play a supportive role in subsequent speech processing.

Taken together, the present study—as part of a larger project—investigated the effects of RAA on the use of MWS and UF in text summary speech and the relationship between them.

Method

Participants

A total of 24 Japanese-speaking learners of English were recruited at a private university in Japan. According to their self-reported scores in English proficiency tests such as TOEFL and IELTS, their proficiency levels ranged from B1 to C1 levels on the CEFR scale.

Materials

Our text summary speaking task included two elements: source texts and their recordings for RAA. We selected two expository texts from Dreamreader.net (Millington, 2015). To maximize the comparability of these two texts, we initially pooled multiple texts and analyzed them in terms of text length, lexical complexity, and readability. Regarding lexical complexity, we used the JACET8000 wordlist which are specifically tailored for Japanese learners of English and replaced vocabulary items above Level 5 with synonyms within Level 1–4. As for readability, we used both a Flesh-Kincaid Reading Ease value based on Coh-Metrix (McNamara et al., 2014) and an overall complexity score based on the TextEvaluator® (Educational Testing Service, 2013) to select a pair

https://doi.org/10.21862/diss-09-009-suzu-korm

of comparable texts. The textual characteristics of the selected texts are summarized in Table 1. The RAA stimuli were recorded by a L1 Canadian English speaker who had 15-year teaching experience of English at universities in Japan. We also ensured the comparability of the delivery speed of recordings across texts (see Table 1).

Table 1. Characteristics of the source texts.

	Text A	Text B
Topic	US Flag	Red Cross
Flesh-Kincaid value	71.21	64.79
TextEvaluator® score	380	660
Common Core Grade	2–3	5–7
Length in words	324	303
Delivery speed (wpm)	116.4	119.6

Procedures

Considering the possibility that individual difference factors may affect the effects of RAA (Liu & Todd, 2014), we decided to use a within-subjects design. In other words, our participants completed both conditions (i.e. [+/− RAA]) while the order of conditions and the source texts were counterbalanced.

For each condition, the participants were first instructed to focus on the gist of meaning of texts rather than the details of information such as dates and were then provided with the source text. The students were allowed to read the texts for approximately three minutes (i.e. the same duration as the RAA recording) either under [+RAA] or [−RAA] conditions. After the text comprehension phase, additional three minutes were given as planning time. During this period, students could plan or rehearse their speech while looking at the source text. Afterwards, participants were instructed to retell the content of the source text in English without looking at the text.

Analysis

The present study analyzed participants' speaking performance in terms of lexical overlaps with source texts and UF. All the audio-recorded speech data were transcribed and annotated for dysfluency phenomena such as self-corrections and false starts. The transcripts were segmented into Analysis of Speech units (AS-units; Foster et al., 2000) as well as clauses and were then submitted to subsequent analyses.

Regarding lexical overlaps, a set of n-gram keyword overlap indices were computed by the Tool for the Automatic Analysis of Cohesion (TAACO 2.0; Crossley et al., 2019): single-words, bigrams, trigrams, and quadgrams. These indices tap into the extent to which characteristic words and n-grams from the source text are used in the speech transcript. TAACO 2.0 identifies single- and multi-word keywords using the news and magazine sections of the Corpus of Contemporary American English (COCA) as a reference corpus (for a detailed description, see Crossley et al., 2019).

As for UF, we employed a set of fluency measures covering speed, breakdown, and repair fluency. Following prior research, we set the duration of silent pauses as 250 milliseconds (Bosker et al., 2013). Using *Praat* software (Boersma & Weenink, 2012), we computed *articulation rate* (AR; the mean number of words per minute, divided by total speech duration excluding pauses), *final-* and *mid-clause pause ratio* (FCPR, MCPR; the mean number of pauses between/within clauses per word), and *dysfluency ratio* (DR; the ratio of dysfluencies to the total number of words).

Results

Lexical overlaps across conditions

In order to detect the differences in lexical overlaps between the source text and subsequent text summary speech across conditions, a set of Wilcoxon signed-ranks tests were performed. As summarized in Table 2, no significant difference was found across conditions except for single words ($Z = 2.25$, $p = 0.024$, $d = 0.36$), suggesting that RAA during text comprehension enhanced lexical overlaps only at the single-word level.

Utterance fluency across conditions

Another set of Wilcoxon signed-ranks tests were performed to examine whether participants' UF significantly differed across conditions. As observed

Table 2. Descriptive Statistics and Wilcoxon signed-ranks tests for single-word and n-gram keywords.

	[– RAA]		[+ RAA]		Wilcoxon-signed rank		
KW% measure	*M*	*SD*	*M*	*SD*	*Z*	*p*	*d*
Single-words	0.115	0.025	0.133	0.041	2.251	0.024	0.355*
Bigrams	0.095	0.057	0.091	0.062	0.200	0.841	0.046
Trigrams	0.034	0.030	0.029	0.033	0.469	0.639	0.097
Quadgrams	0.018	0.021	0.013	0.022	0.719	0.472	0.147

Table 3. Descriptive Statistics and Wilcoxon signed-ranks tests for utterance fluency measure.

		[– RAA]		[+ RAA]		Wilcoxon-signed rank		
Dimension	UF measure	*M*	*SD*	*M*	*SD*	*Z*	*p*	*d*
Speed	AR	104.95	20.84	109.49	20.27	1.743	0.081	–0.397†
Breakdown	FCPR	0.108	0.029	0.114	0.029	0.829	0.407	–0.198
	MCPR	0.404	0.172	0.423	0.179	1.057	0.290	–0.160
Repair	DR	0.193	0.090	0.166	0.099	1.714	0.086	0.343†

Table 4. Correlations between lexical overlap indices and utterance fluency measures

	Single-words		Bigrams		Trigrams		Quadgrams	
UF measure	[–RAA]	[+RAA]	[–RAA]	[+RAA]	[–RAA]	[+RAA]	[–RAA]	[+RAA]
AR	.119	–.301	.079	.104	.250	.438*	.309	.599**
FCPR	–.206	–.020	–.345†	–.543**	–.626**	–.412*	–.535**	–.219
MCPR	.098	.176	–.065	–.267	–.143	–.587**	–.163	–.567**
DR	.048	–.180	.150	–.253	–.174	–.603**	–.299	–.373†

Note. † < .10, * < .05, ** < .01

in Table 3, the results showed that there were only marginally significant changes in AR ($Z = 1.74$, $p = 0.081$, $d = 0.40$) and DR ($Z = 1.71$, $p = 0.086$, $d = 0.34$). In other words, our participants spoke slightly more fluently under the [+RAA] condition during text comprehension.

Interrelationship between lexical overlaps and fluency across conditions

We also examined how RAA affected the inter-relationship between lexical overlaps and UF by calculating Spearman's rank order correlations (see Table 4). Under the [−RAA] condition, the results revealed that only FCPR was correlated with trigram ($rs = −.626$, $p = 0.001$) and quadgram keyword percentages ($rs = −0.533$, $p = 0.007$), indicating that speakers using keyword trigrams and quadgrams tend to produce fewer pauses at clausal boundaries. On the other hand, under the [+RAA] condition, keyword trigrams were correlated with all the UF measures in a supportive direction while keyword quadgrams were supportively correlated with AR ($rs = 0.599$, $p = 0.002$) and MCPR ($rs = −.567$, $p = 0.004$).

Discussion

The effects of read-aloud assistance on lexical overlaps

We found that our L2 speakers used more single-word keywords under the [+RAA] condition than the [−RAA] condition. This positive effect of RAA may indicate that the activation of phonological representations can facilitate the use of single words presented in the source texts. This might also suggest that RAA facilitates text comprehension and that speakers successfully retell the content of the source text using the characteristic words.

The effects of read-aloud assistance on utterance fluency

Meanwhile, we found only marginally significant gains in UF under the [+RAA] condition. L2 speakers tended to produce faster speech with fewer dysfluency phenomena with RAA than without RAA. Regarding speed fluency, the activation of phonological representations may lead to smooth articulatory gestures. Meanwhile, the RAA may reduce two types of self-repairs; the enhanced text comprehension could have reduced information or appropriacy repairs whereas the facilitated selection of lexical items from source texts might have reduced error repairs (Kormos, 2006).

The effects of read-aloud assistance on the interrelationship between lexical overlaps and utterance fluency

The results of correlational analyses revealed that RAA intensified the relationship between n-gram overlaps (mainly, *trigrams* and *quad-grams*) and UF. Although tri- and quad-gram overlaps were correlated only with FCPR under the [−RAA] condition, these two lexical overlap indices were correlated with most of the UF measures under the [+RAA] condition. These correlations under both conditions suggested supportive relationships between multiword lexical overlap and UF measures.

From the perspective of L2 speech production, these results may indicate that the activation of phonological representations, operationalized by

RAA, can facilitate *chunking* during text comprehension (Ellis, 2003). The phonological information such as intonation boundaries can help L2 readers to segment texts into larger units of meaning, and then they can establish the connections among lexical items within the intonation units. As a result, the use of n-gram keywords can positively contribute to UF when L2 learners summarize a text in a dual-mode input condition.

These results may suggest that the activation of phonological representations directly enhances the selection of lexical items and, to a small extent, indirectly facilitates the speed of linguistic encoding processes. In addition, such phonological activation may also play a facilitative role in chunking during text comprehension and in lexical overlaps and fluency of the subsequent text summary speech.

Conclusion

The present study investigated the effects of RAA on the subsequent text summary speech in terms of lexical overlaps and UF in the case of 24 Japanese learners of English. To minimize the effects of individual differences such as preferred modality of input processing, the study used a within-subjects design. The results showed that multi-modal input at the text comprehension phase—reading with RAA—had a direct impact on the proportion of single-word keywords and also an indirect impact on UF in the subsequent text summary speech. In addition, the effects of RAA were found in the interrelationship between multiword lexical overlaps and UF, suggesting that the activation of phonological representations by RAA may facilitate L2 readers' chunking during text comprehension and fluency during subsequent speech.

Finally, it should be noted that the number of participants in our study was relatively small. In relation to the small sample size, we could not investigate the interaction effects between texts and conditions, which we instead tried to minimize by carefully preparing comparable texts. Future research is therefore expected to be conducted with a larger sample size and perform more advanced statistical modelling (e.g. mixed-effects modelling) to carefully examine the relationship between texts, conditions, lexical overlaps and fluency.

References

Boersma, P. & D. Weenink. 2012. Praat: Doing phonetics by computer (version 6.0.46). http://www.praat.org/ (accessed 10 May 2019).

Bosker, H. R., A.-F. Pinget, H. Quené, T. Sanders & N. H. de Jong. 2013. What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, 30(2): 159–175. https://doi.org/10.1177/0265532212455394

Crossley, S. A., K. Kyle & M. Dascalu. 2019. The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, 51(1): 14–27. https://doi.org/10.3758/s13428-018-1142-4

Educational Testing Service. 2013. TextEvaluator®. https://textevaluator.ets.org/ (accessed 10 May 2019).

Ellis, N. 2003. Constructions, chunking and connectionism: The emergence of second language structure. In: C. Doughty & M. H. Long (eds.), *Handbook of second language acquisition*. Malden, MA: Blackwell, 63–103. https://doi.org/10.1002/9780470756492.ch4

Foster, P., A. Tonkyn & G. Wigglesworth. 2000. Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3): 354–375. https://doi.org/10.1093/applin/21.3.354

Kormos, J. 2006. *Speech production and second language acquisition*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Košak-Babuder, M., J. Kormos, M. Ratajczak & K. Pižorn. 2019. The effect of read-aloud assistance on the text comprehension of dyslexic and non-dyslexic English language learners. *Language Testing 36*(1): 51–75. https://doi.org/10.1177/0265532218756946

Liu, Y. T. & A. G. Todd. 2014. Dual-modality input in repeated reading for foreign language learners with different learning styles. Foreign Language Annals, 47(4): 684–706. https://doi.org/10.1111/flan.12113

McNamara, D., A. C. Graesser, P. M. McCarthy & Z. Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

Millington, N. 2015. Dreamreader.net. http://dreamreader.net/ (accessed 10 May 2019).

Skehan, P. 2014. *Processing perspectives on task performance*. Amsterdam: John Benjamins. https://doi.org/10.1075/tblt.5

Stengers, H., F. Boers, A. Housen & J. Eyckmans. 2011. Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? *IRAL — International Review of Applied Linguistics in Language Teaching*, 49(4), 321–343. https://doi.org/10.1515/iral.2011.017

Hesitation patterns in the Spanish spontaneous speech of Hungarian learners of Spanish

Kata Baditzné Pálvölgyi
Department of Spanish Language and Literature, Eötvös Loránd University, Budapest, Hungary

Abstract

This paper examines what native Spanish speakers find most disturbing in the pronunciation of Hungarian language learners of Spanish. Former research (Baditzné Pálvölgyi, 2019) showed that in spontaneous Spanish speech of at least threshold level Hungarian learners, one of the aspects that Spanish native speakers least tolerated was the way Hungarians hesitated. So the present paper focuses primarily on hesitation phenomena—lengthening and filled pauses—assuming that Hungarians hesitate more, and the lengthened segments are longer than the Spanish ones. In order to validate the hypothesis, an investigation comparing a corpus of Northern Spanish spontaneous speech to another corpus of advanced Hungarian learners of Spanish was conducted.

Introduction

The topic of L1 disfluency has been widely discussed in numerous papers, but L2 disfluency phenomena should deserve more attention in research projects (Rieger, 2003) and also in the foreign language classroom (Belz et al., 2017).

According to Medgyes (2001), there are certain areas which, even if language learners stay in the target language country for a long time, they cannot master perfectly. These areas include vocabulary, idiomatic expressions, listening skills, fluency and pronunciation. If we concentrate on these latter two, we find that hesitation is connected to both fields. The sounds people employ when hesitating are not necessarily universal; and if hesitation is defined as the interruption of continuous speech, then it can be seen as a blocker of fluency. On the other hand, hesitation can also guarantee that the speaker will hold the dialogue turn, or at least attract listeners' attention (Bosker et al., 2015), so its role in spoken discourse is twofold. The question of holding turns is of considerable importance in Spanish, as this language is characterized by apparently violent rules concerning debate techniques, which may even cause culture shock to, for example, Hungarians, who are not used to the so-called Mediterranean debate. The ways of holding, and especially obtaining conversational turns seem too vehement for an outsider, and if learners of Spanish are not equipped with the right strategies, they might lose ground in spontaneous discussions with native Spanish speakers, which is obviously not their goal.

Thus we must pay special attention to give learners of Spanish the necessary devices in order to make them self-confident participants in Spanish conversations and debates. As has been mentioned, hesitation schemes can be the clue to assure that the speaker can hold a conversational turn. That is why it is really important that language learners acquire the right hesitation patterns when learning the target language. But do Hungarians learn how to hesitate in Spanish? Even at advanced level, apparently not; earlier research (Baditzné Pálvölgyi, 2019) has shown that Spanish native speakers least tolerated the following aspects in the Spanish spontaneous speech of minimum B1 level Hungarian language learners: the pronunciation of certain vowels and consonants (mostly sibilants); the uncommon intonational patterns employed; the slowness of speech rate, as well as the way Hungarian learners hesitated.

If we examine the differences between Hungarian and Spanish as far as hesitation techniques are concerned, we find that there is a great difference: while Hungarians mostly use the schwa and the [m] to hesitate (Horváth, 2014), the Spanish apply the Spanish vowel [e], as well as [a] and the consonant [m] (Garrido Almiñana et al., 2017). This implies, also based on what native Spanish speakers said about the Spanish pronunciation of Hungarians—i.e. that their speech rate is too slow and they don't hesitate properly—that Hungarians hesitate differently because of mainly two reasons: first, the way they hesitate (tending to use [ə] instead of [e], for instance), and second, the quantity of their hesitation phenomena in their speech as compared to Spanish. This paper focuses on the second problem, formulating the following hypotheses: when hesitating, Hungarian learners of Spanish will:

(1) tend to lengthen segments longer;
(2) tend to hesitate more

than native Spanish people do.

So as to confirm these hypotheses, I carried out a two-step research, first analysing hesitation patterns in a native Spanish corpus, and then contrasting the results with phenomena attested in a corpus provided

https://doi.org/10.21862/diss-09-010-badi

by Hungarian learners of Spanish. The following sections will present the most important findings of the investigation.

The corpus

In order to validate the hypotheses, data was collected from both native Spanish speakers and Hungarian learners of Spanish. The corpus used in this study consisted of audio recordings coming from 25 Northern Spanish speakers (from Asturias, aged 18–23, all learning Humanities at the University of Oviedo), and from 25 Hungarian learners of Spanish (from Budapest, aged 19–25, all learning Spanish Philology at the Eötvös Loránd University). As for the Spanish competence of Hungarian learners, they all reached at least level B1 ('threshold') according to the CEFRL, and had spent no longer than a period of 4 months in Spain (students who had lived in Latin America were excluded from the corpus).

The recordings were conducted by Zoltán Kristóf Gaál (ELTE University), who interviewed the informants within the framework of a Map Task activity (the interviewees had to inform the interviewer about the correct itinerary according to a map). This guaranteed that the recorded speech was spontaneous, and hence suitable for an analysis from the perspective of hesitation phenomena. As the Hungarian informants were not native speakers of Spanish, it took them more time to express themselves and to explain the route. This is why, the total duration of the recordings in case of the language learners was almost double the duration of the native Spanish corpus.

Method

Hesitation is considered to be a disfluency phenomenon. There exist several subcategorizations connected to disfluency phenomena (Neuberger, 2014: 23), but Gósy's (2002) was chosen. She differentiates disfluencies due to erroneous realization from disfluencies due to speaker insecurity. Phenomena related to this latter group include silent pauses, filled pauses (these are properly referred to as hesitations), fillers, repetitions, false starts and lengthenings. As both filled pauses and lengthenings may be considered 'hesitations' in a non-academic terminology, this research focuses on these two disfluency phenomena.

The following aspects were examined in the corpus:

(1) the frequency and duration of lengthened syllables (in case of lexemes), and of filled pauses (in case of vocalized hesitations); and

(2) the proportion of the time dedicated to hesitation phenomena as compared to the total speech time.

Annotation was carried out based on information extracted from the acoustic analysis software Praat (Boersma & Weenink, 2019), bearing in mind the following principle: when measuring lengthening, the syllable was always taken as the basic domain of analysis. This was so in case of the typical Spanish phenomenon of resyllabification (i.e. when a word-final consonant is resyllabified as the onset of the next word starting with a vowel, such as in [la:s alondras], 'the larks'), syllable length was only measured for [la:], as the *s* was resyllabified to the subsequent syllable *a-* in *alondras*.

Results

In order to validate the hypotheses, I examined (a) the percentage of hesitation/lengthening time as compared to the whole speech; (b) the frequency of hesitation/lengthening phenomena per minute and (c) the average duration of hesitations and lengthenings in the case of the 50 informants. Table 1. sums up the results.

Table 1. The characteristics of lengthenings and hesitations in both corpora.

Characteristics of lengthenings and hesitations	Hungarian	Spanish
hesitation and lengthening (%) of the whole speech, mean of 25 speakers	16.93	10.37
Std. dev.	7.64	6.22
Results of *t*-test for Difference of Means (equal variances)	Sig. (1-tailed) 0.00084; $p < 0.05$	
number of lengthening and hesitation phenomena (total)	2092	478
number of lengthening and hesitation phenomena per minute, mean of 25 speakers	20.21	11.22
Std. dev.	6.83	6.51
Results of *t*-test for Difference of Means (equal variances)	Sig. (1-tailed) 0.000009; $p < 0.05$	
duration of hesitation and lengthening phenomena (mean of 25 speakers)	0.51s	0.55s
Std. dev.	0.73	0.82
Results of *t*-test for Equality of Means (equal variances)	Sig. (2-tailed) 0.058; $p > 0.05$	

According to the first hypothesis, Hungarians tend to hesitate more than Spanish do, so I examined the proportion of hesitation phenomena first, that is, the time of lengthening and hesitation produced by

Hungarians compared to their total speech time. Hungarians effectively hesitated in a higher proportion of their total speech time than the Spanish informants did, hesitating during the 16.93% of their whole speech time as opposed to the 10.37% in the Spanish informants' case. These are mean values in case of the 25–25 informants, and statistic testing (one-tailed *t*-test for the difference of means) has revealed that there is a statistical difference between the means at the 95% confidence interval ($p < 0.05$).

This indicates that Hungarians hesitate in a higher proportion of their speech than Spanish do, and this can be due to two factors: the frequency and the duration of their hesitation phenomena. These two aspects were analysed, and as the data show, the difference between Hungarians and Spanish does not lie in the duration but in the frequency of hesitation. Almost four times more hesitation and lengthening phenomena was found in the Hungarian corpus than in the Spanish one. This number itself is impressive but we should bear in mind that Hungarians spoke longer, so the hesitation and lengthening phenomena per minute were analysed in both corpora. The mean values in both corpora were different, so a one-tailed *t*-test was applied again which proved that the difference of the two means—20.21 phenomena per minute in the case of Hungarians as opposed to 11.22 in the Spanish corpus—is statistically different at the 95% confidence interval ($p < 0.05$).

Average duration of hesitation phenomena, on the other hand, seemed no different in the two groups: surprisingly, Hungarians did not produce longer lengthenings compared to native Spanish speakers. The average duration of a hesitation or a lengthening in the case of the Spanish speakers was 0.55 s and 0.51 s in case of the Hungarian speakers, which difference, according to the two-tailed *t*-test for equality of means, is statistically not significant at the 95% confidence interval ($p > 0.05$).

The data confirm the first hypothesis: threshold level Hungarian speakers of Spanish tend to hesitate more than native Spanish speakers, which is due the higher frequency of hesitation phenomena but not to their duration. This also implies that the second hypothesis was not confirmed: segments affected by hesitation phenomena are not considerably longer in case of at least B1 CEFRL level Hungarian learners of Spanish as compared to native Spanish realizations.

Conclusion and discussion

As a reflection on the hypotheses, the data show that, as compared to native Spanish patterns, Hungarian learners of Spanish

(1) do not produce considerably longer lengthened segments than Spanish, but

(2) do hesitate more often.

A further aspect worth considering is a comparison between Hungarian students learning Spanish in foreign language classrooms in a non-immersion context and students learning Spanish abroad. According to the results of García-Amaya (2015) in case of learning Spanish as a foreign language, immersion study context yielded greater fluency and language proficiency improvement but also an increase in the use of filled pauses, for instance.

This raises the question whether the intolerance experimented on behalf of Spanish native speakers when judging Hungarians' Spanish—i.e. that they criticized Hungarian learners' hesitation patterns—was due to the relatively high frequency of hesitation phenomena or rather to other factors, such as the quality (like uncommon vocalic realizations resulting from negative transfer from Hungarian) or the position of hesitation phenomena.

Naturally, the present research is based on data interpretation coming only from Northern Spanish, so in the future further studies are encouraged to analyse hesitation phenomena in other dialects of Spanish as well.

Concerning the duration patterns of speaker-independent hesitation, in the future, it is convenient to stick to an objective model to see how lengthening is realized as compared to its context. Cantero Serena's (2019) duration standardization model would make it possible to see to what extent one lengthened syllable is longer than other syllabic intervals within the same utterance, for instance. Handling relative duration could help us generalize better than if we merely compare absolute duration data.

After defining areas to develop concerning hesitation phenomena, a further question is how to help Hungarian students to improve their hesitation techniques in Spanish in an effective way when no immersion language learning is feasible.

Acknowledgements

Supported by the ÚNKP-19-4 New National Excellence Program of the National Research, Development and Innovation Office. Sincere gratitude also goes to Zoltán Kristóf Gaál, Ph.D. student of the Romance Linguistics Doctoral Programme at Eötvös Loránd University, Budapest, for providing the recordings that were used for setting up the corpora. Also, special thanks to László Pálvölgyi and Péter Pálvölgyi, for their valuable comments on data interpretation.

References

Baditzné Pálvölgyi, K. 2019. Magyar ajkú spanyol nyelvtanulók kiejtése spanyol anyanyelvűek szemével [The pronunciation of Hungarian-speaking Spanish learners through the eyes of native Spanish speakers]. In: *Magyar Alkalmazott Nyelvészeti Kongresszus (MANYE) XXVII*, 15–16 April 2019, Károli Gáspár Reformed University, Budapest, Hungary, 44–44.

Belz, M., S. Sauer, A. Lüdeling & C. Mooshammer. 2017. Fluently disfluent? Pauses and repairs of advanced learners and native speakers of German. *International Journal of Learner Corpus Research* 3(2), 118–148. https://doi.org/10.1075/ijlcr.3.2.02bel

Boersma, P. & D. Weenink. 2019. Praat: Doing phonetics by computer (version 6.0.49). http://www.praat.org/ (accessed 22 March 2019).

Bosker, H. R., J. Tjiong, H. Quené, T. Sanders & N. D. de Jong. 2015. Both native and non-native disfluencies trigger listeners' attention. In: *DiSS 2015, Proceedings of the 7th Workshop on Disfluency in Spontaneous Speech*, 8-9 August 2015, University of Edinburgh, Scotland, UK.

Cantero Serena, F. J. 2019. Análisis prosódico del habla: más allá de la melodía [Prosodic speech analysis: beyond the melody]. In: M. R. Álvarez Silva; A. Muñoz Alvarado & L. Ruiz Miyares (eds.), *Comunicación Social: Lingüística, Medios Masivos, Arte, Etnología, Folclor y otras ciencias afines*. Volumen II, Santiago de Cuba: Ediciones Centro de Lingüística Aplicada, 485–498.

García-Amaya, L. 2015. A longitudinal study of filled pauses and silent pauses in second language speech. In: *DiSS 2015, Proceedings of the 7th Workshop on Disfluency in Spontaneous Speech*, 8-9 August 2015, University of Edinburgh, Scotland, UK.

Garrido Almiñana, J. M., Y. Laplaza & C. L. García. 2017. La caracterización pragmática y prosódica de la vocalización "mmm" en español [Pragmatic and prosodic characterization of the vocalization "mmm" in Spanish]. In: V. Marrero Aguiar & E. Estebas Vilaplana (eds.), *Tendencias actuales en fonética experimental: Cruce de disciplinas en el centenario del Manual de Pronunciación Española (Tomás Navarro Tomás),* Madrid: Uned, 125–129.

Gósy, M. 2002. A megakadásjelenségek eredete a spontán beszéd tervezési folyamatában [The origin of disfluency phenomena in the spontaneous speech planning process]. *Magyar Nyelvőr* 126, 192–204.

Horváth, V. 2014. *Hezitációs jelenségek a magyar beszédben* [Hesitation phenomena in Hungarian speech]. *Beszéd. Kutatás. Alkalmazás.* Budapest: ELTE Eötvös Kiadó.

Medgyes, P. 2001. When the teacher is a non-native speaker. *Teaching pronunciation* 5(12): 429–442.

Neuberger, T. 2014. *A spontán beszéd sajátosságai gyermekkorban* [Characteristics of spontaneous speech in childhood]. *Beszéd. Kutatás. Alkalmazás*. Budapest: ELTE Eötvös Kiadó.

Rieger, C. L. 2003. Disfluencies and hesitation strategies in oral L2 tests. In: R. Eklund (ed.), *Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop*, *Gothenburg Papers in Theoretical Linguistics* 90, 5–8 September 2003 2003, Göteborg University, Sweden, 41–44.

On the role of disfluent speech for uncertainty in articulatory speech synthesis

Charlotte Bellinghausen [1], Thomas Fangmeier [2], Bernhard Schröder [1], Johanna Keller [2], Susanne Drechsel [3], Peter Birkholz [4], Ludger Tebartz van Elst [2] and Andreas Riedel [2]
[1]Institute of German Studies, University of Duisburg-Essen, Essen, Germany
[2]Institute of Psychiatry and Psychotherapy, Medical Center, University of Freiburg, Freiburg, Germany
[3]Department of Speech Science and Phonetics, Martin Luther University Halle-Wittenberg, Halle, Germany
[4]Institute of Acoustics and Speech Communication, TU Dresden, Dresden, Germany

Abstract

In this paper we present a perception study on the role of disfluent speech in forms of prosodic cues of uncertainty in question-answering situations. In our scenario the answer to each question was modeled by varying three prosodic cues: pause, intonation, and hesitation. The utterances were generated by means of an articulatory speech synthesizer. Subjects were asked to rate each answer on a Likert scale with respect to uncertainty, naturalness and understandability. Results showed evidence for an additive principle of the prosodic cues, i.e. the more cues were activated the higher the perceived level of uncertainty. Overall, the effect of intonation and hesitation was more evident than the effect of pause.

Background of the study

The communication of uncertainty

The expression and perception of uncertainty is an essential part in communication (cf. Oh, 2006: 8). In general uncertainty can be regarded as a non-prototypical emotion (Rozin & Cohen, 2003) or as a cognitive state (Kulthau, 1993). We focus on the role of uncertainty in answers following questions. For the acoustic channel several studies suggested evidence that uncertainty is not only expressed, but also perceived by prosodic means like rising intonation, pauses, and hesitations (Smith & Clark, 1993; Brennan & Williams, 1995; Swerts & Krahmer, 2005).

With respect to disfluent speech in acoustic speech synthesis, the synthesis of filled pauses (Adell, Bonafonte & Escudero-Mancebo, 2010) and also of filled pauses and hesitations (Andersson et al., 2010) in Unit Selection Synthesis does not show an increase of naturalness. In Hönemann and Wagner (2016) uncertainty is modelled as one of four emotional states by means of prosodic and voice quality parameters. Furthermore, decreased vocal effort, filled pauses and prolongation of function words contribute to uncertainty perception in synthetic speech using a corpus based-method (Székely, Mendelson & Gustafson, 2017).

Articulatory speech synthesis

In our approach we used the articulatory synthesizer VocalTractLab (Birkholz, 2017), which allows to generate high quality speech sounds while manipulating parameters of the time varying laryngeal and supra-laryngeal actions. The synthesizer consists of a geometric 3D model of a male vocal tract (Birkholz, 2013) controlled by 23 parameters to simulate the articulation. The voice source is generated by a self-oscillating model of the vocal folds (Birkholz, Kröger & Neuschaefer-Rube, 2011) which is controlled by six parameters to specify the subglottal pressure, fundamental frequency, and the rest shape of the glottis. The movements of the 3D model and the fundamental frequency are controlled by a gestural score. For each synthetic word the articulatory movements are adjusted manually and generated with different prosodic features.

Previous work

In our previous work (Lasarcyk et al., 2013; Wollermann et al., 2013) we investigated perceived uncertainty by using prosodic cues. The stimuli were question-answer pairs in a human-machine scenario. The answer varied with respect to the combination of the cues pause (absent vs. present), intonation (falling vs. rising) and hesitation (absent vs. present). The experiment design was characterized by three blocks, each time with a 2 × 2 design with pause vs. intonation, intonation vs. hesitation and hesitation vs. pause as independent variables.

261 students of University of Duisburg-Essen (all native speakers of German) listened to the question-answer pairs. They had to rate each time on a 5-point Likert scale how uncertain the answer sounds, how natural, and how understandable it sounded. Results showed in general that the cues of uncertainty were additive with respect to perceived uncertainty.

Perception of uncertainty in ASD

In our interdisciplinary project we investigate the perception of prosodic indicators of uncertainty in Autism Spectrum Disorder (ASD). The aim of our

https://doi.org/10.21862/diss-09-011-bell-etal

current perception study (see below) is to validate the material by presenting it to neurotypically developed subjects. According to diagnostic criteria of DSM-5 (Falkai & Wittchen, 2015) ASD is a neurodevelopmental disease characterized by difficulties in social communication, unusually restricted, repetitive behavior and interests, and specific differences in language and perception. It is mainly accompanied by qualitative deviations in mutual interactions and patterns of communication.

There is an increasing number of studies investigating the role of prosody in ASD. Diehl and Paul (2011) found differences between the perception and imitation of prosodic patterns in children with ASD compared to the control group. Furthermore, in the context of perceiving information status adult listeners with Asperger Syndrome made less use of prosody than the control group, they, however, rely more on lexical information like word frequency and semantic information (Grice, Krüger & Vogeley, 2016).

With respect to emotion perception in articulatory speech synthesis, Hsu and Xsu (2014) showed that high-functioning autistic listeners were less sensitive with respect to emotional prosody by manipulating voice quality as compared to the control group.

Perception study

Goal

The following questions were addressed: Are subjects able to discriminate different intended levels of uncertainty expressed by the three prosodic cues pause, intonation, and hesitation? Is there a correlation between the judgments of uncertainty and the judgments of naturalness as well as of understandability?

Material

Our stimuli were question-answer pairs between a research assistant and a robot for image recognition which were part of a human-machine scenario. The assistant showed pictures of fruits and vegetables to the robot and asked the robot "Was siehst Du?" / *What do you see?* The robot recognized the objects with a certain confidence score, depending on the quality of the picture. Thus, the system was able to express uncertainty about recognition in its answer which was a one-word sentence. As our critical stimuli we chose four one-word trisyllabic phrases in German: "Bananen", "Limetten", "Melonen", "Tomaten" / *bananas*, *limes*, *melons*, and *tomatoes*. For each critical stimulus nine different intended levels of uncertainty were generated (see Table 1).

a) **Pause** refers to the time between the question and the answer. For every level of intended uncertainty we used a default silence pause of 1 s. In the case of pause[+] we used either a silent pause of 4 s or a filled pause, i.e. the hesitation "äh" / uh which took 0.37 s followed by a silent pause of 3.632 s, such that the total duration of the whole pause was 4 s. Since it was not clear from the literature which length of pause is exactly adequate for our research question, we chose an obviously marked pause of 4 s to see whether there is any effect at all on uncertainty perception. b) As **hesitation** particle we chose the particle "äh" / uh since this particle occurred most often for the Verbmobil corpus in German (Batliner et al., 1995). It was either activated (hes[+]) or deactivated (hes[−]). c) The **intonation** of the intended level of certainty showed a peak (measured in semitones) on the stressed syllable of the word with 37 ST. To express uncertainty the last syllable was either characterized by 38 ST for slight uncertainty (level of uncertainty 1) and by 44 ST for strong uncertainty (level of uncertainty 2). Figure 1 shows the different intonation contours for the critical stimulus "Bananen" (each time the question is preceding).

In addition to the critical stimuli, we used nine further one-word phrases as distractors. The utterances were "Birnen", "Blaubeeren", "Bohnen", "Erdbeeren", "Gurken", "Knoblauch", "Mandarinen", "Orangen", and "Paprika" / *pears*, *blueberries*, *beans*, *strawberries*, *cucumbers*, *garlic*, *mandarins*, *oranges*, and *paprika*. The distractors were all characterized by the absence of all three uncertainty cues. We used them in order to distract subjects from the critical stimuli.

Hypothesis

Based on our previous findings (see "Previous work" above) we expected that the prosodic cues of uncertainty have an additive effect, i.e. the activation of all three cues yields a higher degree of perceived uncertainty.

Design

In total we used 36 critical stimuli (4 critical stimuli × 9 different levels of intended uncertainty), 9 stimuli as distractors and one example stimulus. In order to minimize learning effects of the subjects we divided the stimuli into four subsets, each with 19 question-answer pairs in a different random order.

Table 1. Nine different levels of intended uncertainty

pause	hesitation	intonation	level
-	-	-	certainty (c)
-	+	-	hesitation (hes)
+	-	-	pause (pau)
-	-	+	intonation 1 (into 1)
-	-	+	intonation 2 (into 2)
+	+	-	hes & pau
-	+	+	hes & into 2
+	-	+	pau & into
+	+	+	pau & into & hes

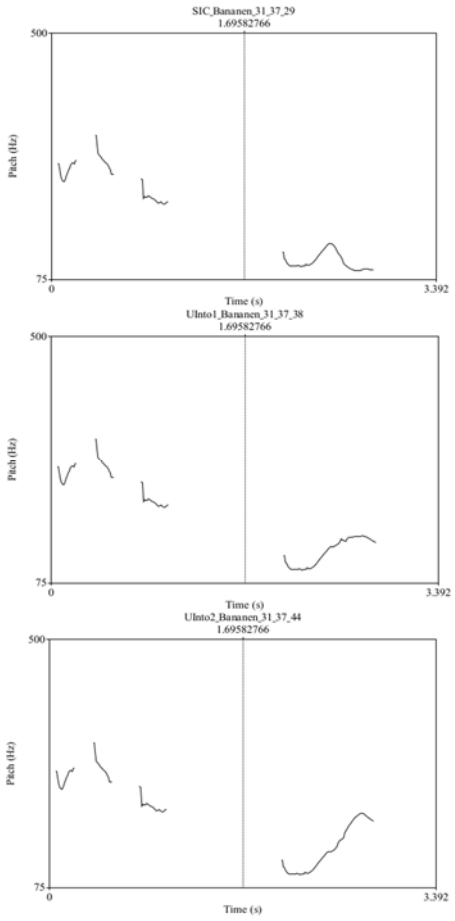


Figure 1. Intonation contour for intended level of a) pattern 31-37-29 for certainty (top), b) pattern 31-37-38 for intonation 1 (middle), c) pattern 31-37-44 for intonation 2 (bottom).

Procedure

Subjects were 36 undergraduate students (23f, 13m; average age: 25 years) of Duisburg-Essen University. All of them were native speakers of German. The number of students per group was as follows: G1: $N = 10$; G2: $N = 7$; G3: $N = 9$; G4: $N = 10$. The procedure started with the presentation of the example stimuli in a seminar room. After each of the 19 question-answer pairs was played subjects had to rate on three 5-point Likert scales a) how uncertain the answer of the robot sounded, b) how natural it sounded, and c) how understandable it was. In addition, subjects had to list the word which they perceived in the answer.

Statistical analysis

For making comparisons between the rankings of the different levels of uncertainty we performed Wilcoxon Matched Pairs Tests with Bonferroni correction. Since we had 30 comparisons our alpha was 0.05/30=0.00167. Furthermore, we tested by means of Spearman Rho Test whether there was a correlation between uncertainty perception and perception of a) naturalness and b) understandability.

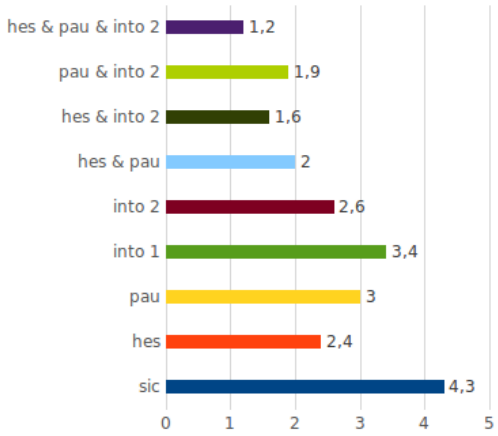


Figure 2. Judgments for perceived uncertainty; x-axis: mean; y-axis: intended level of uncertainty, pau: pause, hes: hesitation, into 2: intonation 2 (see also Figure 1)

Table 2. Significance values of pairwise comparisons using Wilcoxon Matched Paired Test. Significant results with $p < 0.00167$ are marked by x.

	sic	hes	pau	into 1	into 2
hes	x				
pau	x	n.s.			
into 1	x	x	n.s.		
into 2	x	n.s.	n.s.	n.s.	
hes & pau	x	n.s.	x		n.s.
hes & into 2	x	x	x		x
pau & into 2	x	n.s.	x		x
hes & pau & into 2	x	x	x	x	x

	hes & pau & into 2
hes & pau	n.s.
pau & into 2	x
hes & into 2	x

Results

The results for the perceived uncertainty are shown in Figure 2. In Table 2 results for the pairwise comparisons are shown.

The level of intended certainty was ranked significantly different from all other levels of intended uncertainty. When only one prosodic cue was activated the perceived level of uncertainty was always lower in a significant way as opposed to the activation of all three cues. Comparing the activation of a single cue with the activation of two cues, the additional pause combined with hesitation did not contribute significantly to perceived uncertainty. For the other cases, the additional effect was significant. When two activated cues of intended uncertainty are compared to three activated cues the results are as follows: pause and hesitation have an additive effect on the perceived uncertainty in a significant way, but intonation 2 does not. With respect to the comparisons between single cues, our data showed in general no significant differences between judgments except for hesitation vs. intonation 1. For the correlation between the judgment of uncertainty and a) naturalnesss and

also of b) understandability we had 10 calculations such that our a was .01/10 = 0.001. The Spearman's Rho Test computed for case a) $p = 0.692$ and for case b) $p = 0.003$. Thus, no significant correlation was found.

Discussion

We presented a study investigating the role of prosodic indicators for uncertainty perception. Results in general suggest an additive effect of prosodic cues and are in line with our previous findings (Lasarcyk et al., 2013; Wollermann et al., 2013). The relative contribution of the pause to uncertainty perception is not clear from our data. For future work we would like to test in a more fine-grained way the effect of pause length. The current study shows that the tested prosodic features of intended uncertainty provide us with a sufficient number of degrees of uncertainty as perceived by neurotypical subjects, so that the stimulus material can be considered suitable to test whether there are differences between neurotypical and autistic hearers.

Acknowledgments

The project is funded by the "Programm zur Förderung des exzellenten wissenschaftlichen Nachwuses" of University of Duisburg-Essen.

References

Adell, J., A. Bonafonte & D. Escudero-Mancebo. 2010. Modelling Filled Pauses Prosody to Synthesize Disfluent Speech. In: *Proceedings of Speech Prosody 2010*, 10–14 May 2010, Chicago, IL, 100624, 1–4.

Andersson, S., K. Georgila, D. Traum, D., M. Aylett & R. A. J. Clark. 2019. Prediction and Realisation of Conversational Characteristics by Utilising Spontaneous Speech. In: *Proceedings of Speech Prosody 2010*, 10–14 May 2010, Chicago, IL, 100116, 1–4.

Batliner, A., A. Kieling, S. Burger, & E. Nöth. 1995. Filled Pauses In Spontaneous Speech. In: *Proceedings of 13th International Congress of Phonetic Sciences (ICPhS)*, 14–19 August 1995, Stockholm, Sweden, vol. 3, 472–475.

Birkholz, P. 2013. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS ONE* 8:e60603. https://doi.org/10.1371/journal.pone.0060603

Birkholz, P. 2017. Vocal Tract Lab (version 2.2). http://www.vocaltractlab.de/ (accessed 03.09.2019).

Birkholz, P., B. J. Kröger, C. Neuschaefer-Rube. 2011. Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis. In: *Proceedings of Interspeech*, 27–31 August 2011, Florence, Italy, 2681–2684.

Brennan, S. E. & M. Williams, M. 1995. The feeling of anothers knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language* 34, 383–398. https://doi.org/10.1006/jmla.1995.1017

Diehl, J. & R. Paul. 2011. Acoustic differences in the imitation of prosodic patterns in children with autism spectrum disorder. *Research in Autism Spectrum Disorders* 6(1): 123–134. https://doi.org/10.1016/j.rasd.2011.03.012

Falkai, P. & H.-U. Wittchen (eds.). 2015. *Diagnostisches und statistisches Manual psychischer Störungen: DSM-5* [Diagnostic and statistical manual of mental disorders: DSM-5]. Göttingen: Hogrefe, 64ff.

Grice, M., M. Krüger & K. Vogeley. 2016. Adults with Asperger syndrome are less sensitive to intonation than control persons when listening to speech, *Culture and Brain* 4(1): 38–50. https://doi.org/10.1007/s40167-016-0035-6

Hönemann, A. & P. Wagner. 2016. *Synthesizing Attitudes in German"*. In: *Proceedings of the 16th Speech Science and Technology Conference*, 6–9 December 2016, Sydney, Australia, 209–213.

Hsu, C. & Y. Xu. 2014. Can adolescents with autism perceive emotional prosody? In: *Proceedings of Interspeech 2014*, 14–18 September, Singapore, 1924–1928.

Kuhlthau, C. C. 1993. *Seeking Meaning: A Process Approach to Library and Information Services*, Norwood, NJ: Ablex.

Lasarcyk, E., C. Wollermann, B. Schröder & U. Schade. 2013. On the Modelling of Prosodic Cues in Synthetic Speech – What are the Effects on Perceived Uncertainty and Naturalness? In: *Proceedings of the 10th International Workshop on Natural Language Processing and Cognitive Science*, 15–16 October 2013, Marseille, France, 117–128.

Oh, I. 2006. *Modeling Believable Human-Computer Interaction with an Embodied Conversational Agent: Face-to-Face Communication of Uncertainty*, Rutgers The State University, Dissertation.

Rozin, P. & A. B. Cohen. 2003. High Frequency of Facial Expressions Corresponding to Confusion, Concentration, and Worry in an Analysis of Naturally Occurring Facial Expressions of Americans. *Emotion* 3(1): 68–75. https://doi.org/10.1037/1528-3542.3.1.68

Smith, V. L. & H. H. Clark. 1993. On the Course of Answering Questions. *Journal of Memory and Language* 32(1), 25–38. https://doi.org/10.1006/jmla.1993.1002

Swerts, M. & E. Krahmer. 2005. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language* 53(1), 81–94. https://doi.org/10.1016/j.jml.2005.02.003

Szekely, E., J. Mendelson & J. Gustafson. 2017. Synthesising uncertainty: The interplay of vocal effort and hesitation disfluencies. In: P*roceedings of Interspeech*, 20–24 August 2017, Stockholm, Sweden, 804–808. https://doi.org/10.21437/Interspeech.2017-1507

Wollermann, C., E. Lasarcyk, U. Schade & B. Schröder. 2013. Disfluencies and Uncertainty Perception – Evidence from a Human-Machine Scenario. In: R. Eklund (ed.): *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech*, 21–23 August 2013, Stockholm, Sweden, 73–76.

"Uh" is preferred by male speakers in informal presentations in American English

Michiko Watanabe [1], Yusaku Korematsu [2] and Yuma Shirahata [2]
[1]Center for Corpus Development, National Institute for Japanese Language and Linguistics, Tokyo, Japan
[2]School of Engineering, The University of Tokyo, Tokyo, Japan

Abstract

This study investigates factors that are likely to be related to speakers' choice of filler type between uh and um in English, using an informal presentation speech corpus. The effects of the following factors on the probability of each filler type was examined: (1) immediately preceding clause boundary depth, (2) clause size measured as the number of words in the clause, (3) the number of quotation remarks in the clause, and (4) speaker's sex. The filler probabilities increased with the boundary depths. This trend was much stronger with um than with uh. Ums are more likely to appear clause-initially than uhs. Clause size had similar effect sizes on the two filler types. The number of quotation remarks had a stronger negative effect with ums. Speaker's sex had a significant effect only with uhs. Uhs are used more frequently by male speakers than by female speakers. The results indicate that speakers' choice of filler type is affected by the combination of multiple factors with various effect sizes.

Introduction

When speakers need extra time to continue their speech, they are likely to slow down the current speech, pause longer, repeat syllables or words, and/or utter fillers to buy time for speech planning. *Uhs* and *ums* are the most common filler types in English. Are there any difference in the usage or functions between these two types? This study investigates factors that are likely to affect speakers' choice of fillers between the two types.

Clark and Fox Tree (2002) found that *uh* and *um* differ with respect to the length of following silent pauses. Long pauses are more likely to be preceded by *um* than by *uh*. From this finding, the authors argue that filler type informs listeners of the seriousness of upcoming problems. On the other hand, Brennan and Williams (1995) observed no difference in listeners' impression about the metacognitive states of speakers between the two types.

It was observed that *ums* are more frequent at deeper discourse boundaries than *uhs* in Dutch monologues (Swerts, 1998). Shriberg (1994) observed that *um*'s rate is higher sentence-initially than sentence-medially, whereas *uh*'s rate is higher sentence-medially than sentence-initially in "The Switchboard Corpus", a corpus of informal telephone conversations in American English. These studies suggest that boundary depth is related to speaker's choice between the two types. *Um* has been reported to be more common at deeper boundaries than *uh*. We first examine whether *um* is more likely to appear clause-initially than *uh* by examining the ratios of *ums* and *uhs* at clause-initial and clause-medial positions in "The Corpus of Oral Presentations in English" (COPE). The details of the corpus are given in the method section.

Tottie (2011) observed that male speakers use *uhs* more frequently than *ums* whereas female speakers use roughly the same number of *uhs* and *ums* in British English. It was also observed that *ums* are more frequently used by younger speakers and by those with higher socio-economic status in British English. On the other hand, no sex difference or clear socio-economic effect was observed in American English (Tottie, 2014). These studies indicate that sociolinguistic factors affect speaker's choice of filler type. We examine sex effect on the probability of *uh* and *um* in the present study.

Watanabe and Korematsu (2019) investigated factors that are likely to be relevant to the clause-initial filler probability, using COPE. It was observed that the clause size—measured as the number of words in the clause—as well as the preceding clause boundary depth are related to the clause-initial filler probability. The filler probability increased with the clause boundary depth and with the number of words in the clause. It was also found that the number of quotation remarks in the clause negatively affects the clause-initial filler probability. Sex was also examined as a factor, but no significant effect was observed. Based on these findings, we included the clause size and the number of quotation remarks in the clause as possible predicting factors of the probability of *uh* and *um* at clause-initial and clause-medial positions in the present study.

Method

Speech material

We used COPE as material. COPE contains twenty informal English presentation speeches recorded in

https://doi.org/10.21862/diss-09-012-wata-etal

Portland, Oregon and Los Angeles, California in 2013. An overview of COPE is given in Table 1.

Table 1. Overview of COPE.

Number of speakers	20
Total duration (minutes)	227
Total number of words	38,370
Total number of fillers	1,441
Clause initial fillers	941
Clause medial fillers	500

The speakers were university students or university graduates in their 20s and early 30s. They were given a topic, "the most memorable event in my life", and instructed to give a talk for at least 10 minutes. They gave their talks in front of a small audience including their friends and the recording staff. Disfluencies were labeled, and clause boundaries were marked on the transcription. Only *uh* and *um* were identified as fillers. An excerpt from the transcription of COPE is given in Example (1) below.

Example (1)
1: <cb and when> and when we got to (l like) (r2 this little) this little food court area <ce and when>
2: /cb/ he stopped me /ce/
3: /cb and/ and he was like (qcb) so would you want (ncb to) to be my girlfriend (nce to) (qce) /ce and/
4: /cb and/ and I literally jumped on him /ce and/
5: /cb and/ and I have no idea (ncb why) why I did that (nce why) /ce and/
(From losF09 in COPE)

Transcriptions of COPE are partitioned at coordinate or adverbial finite clause boundaries as in Example (1). Ankle brackets, <cb> and <ce> indicate finite adverbial clause beginning and end, respectively. Connectives are included within the brackets. "/cb/" and "/ce/" in line 2 mean coordinate clause beginning and end, respectively. "(qcb)" and "(qce)" in the third line indicate quotation clause beginning and end, respectively. "(ncb to)" and "(nce to)" in the third line indicate nominal infinitive clause beginning and end, respectively. Similarly, "(ncb why)" and "(nce why)" in the fifth line indicate nominal clause beginning with "why" and its end, respectively.

Procedure

The following factors were examined as those possible to predict the probability of *uh* and *um*.
(1) Clause boundary depth
Whether and how the effect size of clause boundary depth differs between *uh* and *um* was examined.

In English, it is difficult to tell whether a given clause boundary is also a sentence boundary or a boundary within a sentence, because no period or comma is used in speech. There is no syntactic or morphological cue to tell boundary depth in English. In order to evaluate clause boundary depth, we had three labellers add sentence boundary labels to clause boundaries which they judged to be deep boundaries. The labellers were instructed to mark sentence boundaries based on the content and the prosody of speech. We estimated boundary depth based on the number of labellers who marked the boundary as a sentence boundary. We regarded boundaries marked by three labellers as sentence boundaries to be the deepest, those marked by two labellers to be the second deepest, and so forth. Thus, clause boundaries were grouped into four types depending on perceived boundary depth. The number of boundaries in each group is given in Table 2.

Table 2. Number of boundaries in each group.

Type0	Boundaries marked by no labeler	719
Type1	Boundaries marked by one labeler	1521
Type2	Boundaries marked by two labelers	979
Type3	Boundaries marked by three labelers	806
	Total	4025

(2) Clause size
The number of words in each clause was counted and regarded as an index of message complexity to be conveyed by the clause. Fillers and other disfluencies were not counted as words.
(3) Number of quotation clauses
Speech sometimes contains quoted remarks. The number of quotation clauses in a given clause was found to have a negative effect on the containing clause initial filler probability (Watanabe and Korematsu, 2019). The number of quotation clauses was included as a possible predicting factor of the probability of *uh* and *um*.
(4) Speaker's sex
Speaker's sex was included as a predicting factor, because Tottie (2011, 2014) reports that male speakers use more *uhs* than *ums* while female speakers use roughly the same number of the two types in British English, whereas no such difference was observed in American English.

We excluded clauses containing more than 30 words from analysis because samples were sparse for larger clauses.

We estimated the probability of *uh* and *um* at clause-initial and clause-medial positions separately using a generalized linear mixed model, with maximum likelihood estimation of variance components. Because the response variable was

binary, we conducted mixed-effects logistic regression. Fixed effects predictor variables were the factors (1) through (4). Speakers were treated as a random effects factor. We used `glmer` function in `lme4` package and `MuMIn` and `lmerTest` packages running under R version 3.5.1.

Results

The proportion of clause-initial and clause-medial uh and um

Table 3 shows the frequency and the ratio of *uh* and *um* at clause-initial and at clause-medial positions. *Ums* are far more frequent than *uhs* clause-initially, whereas the difference is small clause-internally. The ratio shows that 73% of *ums* are used clause-initially, while *uhs* are used roughly equally clause-initially and clause-medially. The results are in accordance with the results of previous studies that *ums* tend to be used at deeper boundaries than *uhs*.

Table 3. Frequencies and ratios of ums and uhs at clause-initial and clause-medial positions.

	Frequency			Ratio	
	Initial	Medial	Total	Initial	Medial
uh	245	259	504	0.49	0.51
um	683	254	937	0.73	0.27

Factors related to clause-initial uhs and ums

Table 4 shows the model results of clause-initial *uhs*, and Table 5 *ums*. Estimates of categorical variables are given as the relative values to one of its levels whose estimate is zero in R. The reference variables are included in the table. Odds ratio indicates the degree of effect size of each factor. When the ratio is close to 1.0, the effect size is small. The more distant the ratio from 1.0, the larger the effect size.

First, both clause-initial *uhs* and *ums* are significantly related to the boundary depth. The filler probabilities increase with the boundary depths. This trend is much stronger with *um* than with *uh*. Second, the clause size has a significant effect on both filler types. The effect sizes are very close to each other. Third, the number of quotation clauses has a significant negative effect only with *um*. *Uh* is not affected by the factor. Fourth, sex effect is significant only with *uh*. Male speakers use significantly more *uhs* than their female counterparts.

Table 4. Results of mixed-effects logistic regression for clause-initial uh.

Variable		Estimate	Std. Error	z value	Pr(>|z|)		Odds ratio
(Intercept)		-5.068	0.3868	-13.103	< 2e-16	***	0.006
Boundary depth		0.398	0.070	5.667	0.000	***	1.488
Number of words		0.039	0.012	3.176	1.49e-03	**	1.040
Number of quotation clauses		-0.092	0.146	-0.633	5.26e-01		0.912
Gender	Female	0					
	Male	1.655	0.456	3.628	0.000	***	5.233

Table 5. Results of mixed-effects logistic regression for clause-initial um

Variable		Estimate	Std. Error	z value	Pr(>|z|)		Odds ratio
(Intercept)		-3.966	0.326	-12.184	< 2e-16	***	0.019
Boundary depth		1.059	0.054	19.440	< 2e-16	***	2.884
Number of words		0.045	0.009	4.794	1.63e-06	***	1.046
Number of quotation clauses		-0.708	0.184	-3.846	0.000	***	0.493
Gender	Female	0					
	Male	-0.192	0.416	-0.462	0.644		0.825

We reanalyzed 90% of the data only with significant variables with speakers as a random effects factor, and estimated filler probability for the remaining data.

Figure 1 and Figure 2 illustrate estimated probabilities of clause-initial *uh* and *um* for each boundary type as a function of the number of words in the clause. As sex factor had a significant effect on uh, the probability of *uh* is shown separately for female and male speakers. Filled circles indicate observed boundaries with fillers and pluses indicate

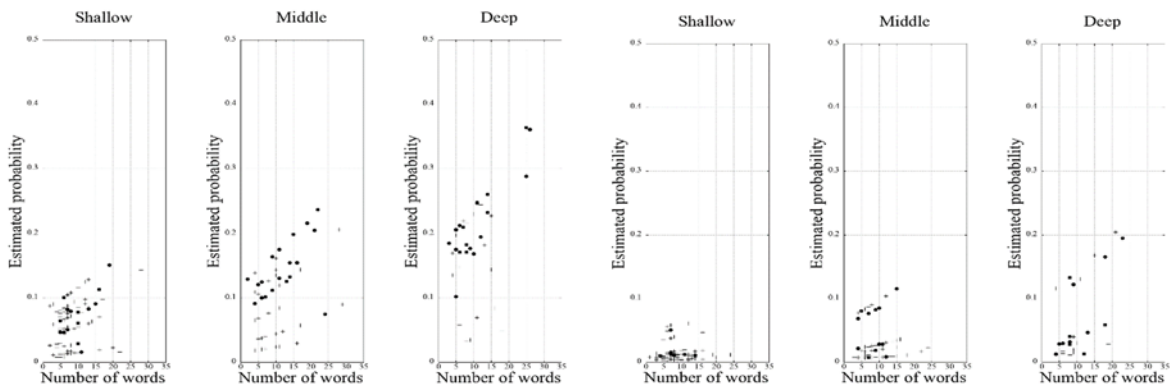


Figure 1. Estimated clause-initial probability of uh for each boundary type as a function of number of words in the clause: the upper figure for male speakers and the lower for female speakers. Filled circles indicate observed boundaries with fillers and pluses without fillers.

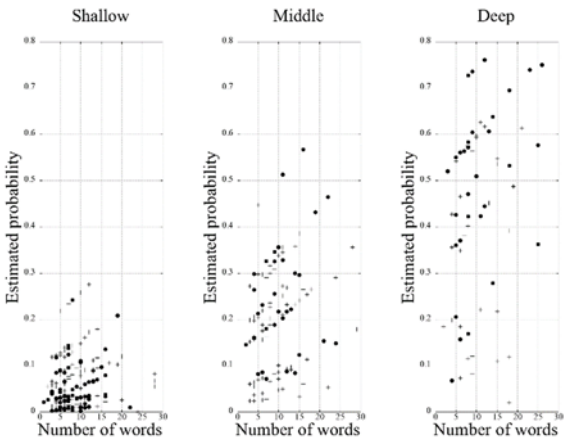


Figure 2. Estimated clause-initial probability of um for each boundary type as a function of number of words. Filled circles indicate observed boundaries with fillers and pluses without fillers.

observed boundaries without fillers. Type0 and Type1 boundaries in Table 2 are shown combined as shallow boundaries.

A comparison of Figure 1 and Figure 2 demonstrates difference in effect size of boundary depth. A comparison of the effects on female and male speakers in Figure 1 indicates that sex effect is not small with *uh*. Male speakers use *uh* at a much higher rate than female speakers.

Factors related to clause-medial ums and uhs

Table 6 shows the results of clause-medial *uhs*, and Table 7 *ums*. Boundary effects with clause-medial fillers are weaker than those with clause-initial fillers for both *uh* and *um*. The effects are marginally significant for both types. Effect size of number of words with clause-medial fillers are larger than that with clause-initial fillers for both types. Effects of number of quotation clauses are significant for both types. Sex effect is significant only with *uh*, as is the case with clause-initial fillers.

Discussion

The results are in accordance with those of previous research that *um* is preferred clause-initially than *uh* (Shriberg, 1994; Swerts, 1989). Message units between deep boundaries are generally larger than those between shallow boundaries. It is inferred that *um* reflects speaker's cognitive load of planning larger units than *uh*.

Clause size measured as the number of words had similar effect sizes on *uh* and *um*. It is likely that *uh* and *um* reflect the message complexity to be conveyed in the clause to a similar degree.

Table 6. Results of mixed-effects logistic regression for clause-medial uh.

Variable		Estimate	Std. Error	z value	Pr(>|z|)		Odds ratio
(Intercept)		-5.594	0.3829	-14.611	<2e-16	***	0.004
Boundary depth		0.133	0.076	1.759	0.079	.	1.142
Number of words		0.133	0.012	10.946	<2e-16	***	1.142
Number of quotation clauses		-0.288	0.146	-1.979	4.78e-02	*	0.750
Gender	Female	0					
	Male	1.341	0.447	3	0.003	**	3.823

Table 7. Results of mixed-effects logistic regression for clause-medial um.

Variable		Estimate	Std. Error	z value	Pr(>|z|)		Odds ratio
(Intercept)		-5.261	0.334	-15.75	< 2e-16	***	0.005
Boundary depth		0.153	0.078	1.959	0.050	.	1.165
Number of words		0.185	0.013	14.531	< 2e-16	***	1.203
Number of quotation clauses		-0.687	0.198	-3.472	0.001	***	0.503
Gender	Female	0					
	Male	-0.368	0.383	-0.962	0.336		0.692

A sex effect was observed with *uh*. It is likely that sociolinguistic factors play a role in the choice of filler type not only in British English but also in American English.

Acknowledgements

This research is supported by JSPS KAKENHI, Grant Numbers 15K02553 and 18K00559.

References

Brennan, S. E. & M. Williams. 1995. The feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and language* 34(3): 383−398. https://doi.org/10.1006/jmla.1995.1017

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

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

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

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

Tottie, G. 2011. Uh and Um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics* 16(2): 173−197. https://doi.org/10.1075/ijcl.16.2.02tot

Watanabe, M. & Y. Korematsu. 2019. Comparison of factors related to clause-initial filler probabilities in English and Japanese, *Proceedings of ICPhS 2019, the International Congress of Phonetic Sciences*, 4–10 August 2019, Melbourne, Australia, 2440–2444.

Segment prolongation in Hebrew

Vered Silber-Varod [1], Mária Gósy [2] and Robert Eklund [3]
[1]Open Media and Information Lab (OMILab), The Open University of Israel, Israel
[2]Dept. of Phonetics, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
[3]Dept. of Culture and Communication, Linköping University, Sweden

Abstract

In this paper we study segment prolongations (PRs), a type of disfluency sometimes included under the term "hesitation disfluencies", in Hebrew. PRs have previously been studied in a number of other languages within a comprehensive speech disfluency framework, which is applied to Hebrew in the current study. For the purpose of this study we defined Hebrew clitics, such as conjunctions, articles, prepositions and so on, as words. The most striking difference between Hebrew and the previously studies languages is how restricted PRs seem to be in Hebrew, occurring almost exclusively on word-final vowels. The most frequently prolonged vowel is [e]. The segment type does not affect PRs' duration. We found significant differences between men and women regarding the frequency of PRs.

Introduction

Prolongation (PR), long being studied in stuttered speech as a characteristic phenomenon, was found to be very common in non-pathological speech, as a disfluency phenomenon that is used when a speaker is hesitating (Eklund & Shriberg, 1998; Eklund, 2001, 2004: 163), i.e. when a syllable or 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, PRs have been shown to differ from FPs in some respects (e.g. Eklund, 2001). Previous studies of PRs have covered a wide variety of languages that are taxonomically similar, such as English and Swedish (Eklund & Shriberg, 1998), or taxonomically different, such Swedish and Tok Pisin (Eklund, 2001, 2004), Japanese (Den, 2003), and Mandarin Chinese (Lee et al, 2004). Other languages studied include German (Betz, Eklund & Wagner, 2017), and Hungarian (Gósy & Eklund, 2018).

One characteristic of PRs that have been discussed in those studies is how they are distributed in the word in which they appear. Simply put, a segment subjected to prolongation (above normal duration, relative to speech rate), can appear in three basic positions:

- Word *initial*, e.g. "ssssegment" [s:egment]
- Word *medial*, e.g. "segmmmmment" [segm:ent]
- Word *final*, e.g. "segmen t" [segment:]

Eklund and Shriberg (1998) reported for American English and Swedish, a 30–20–50% distribution, for initial–medial–final positions, respectively. What made these figures interesting, however, were studies of other languages. Eklund (2001; 2004: 251) reported 15–0–85% in Tok Pisin; Den (2003) reported 0–5–95% for Japanese; and Lee et al. (2004) reported 4–1–95% for Mandarin Chinese.

The interesting difference was that Japanese, Mandarin and Tok Pisin exhibit both much less complex syllable structures, and even less complex phonotactics, as compared to English and Swedish. This led Eklund (2004: 251) to coin the 'morphology matters hypothesis', although Eklund later rephrased this (somewhat misleading term) as 'phonotactics matters hypothesis' instead (Gósy & Eklund, 2018). What is of interest here is that the relative distribution of PRs seems to be, at least partly, depending on the underlying phonotactics of the language in question, where more permissive phonotactics result in a more even PR distribution in the word. The reason for this change was that Betz, Eklund and Wagner (2017) reported 7–15–78% for German, and Gósy and Eklund (2018) reported 18–19–63 for Hungarian, thus making the overall impression slightly more muddled than the impression given by the previously mentioned studies; for example, German, although more similar to English and Swedish from a morphological and phonotactical point of view, lies closer to Mandarin when it comes to initial PRs, and Hungarian lies closer to Tok Pisin for initial position, than either language does when compared to English or Swedish.

In Hebrew, PRs and FPs were thoroughly studied (Silber-Varod, 2013a; 2013b), however with the current comprehensive speech disfluency framework, the Hebrew PR phenomenon would be comparable to other languages. Hebrew is a Semitic language that belongs to the group of Afro-Asiatic languages, taxonomically different from the previously studied languages. Although classical Hebrew is a fusional language (also called inflected language), with a rich morphology and an extensive system of affixation, Modern Hebrew is considered more analytic compared to rather synthetic Classical Hebrew. The inflected characteristics of Hebrew have implications on the syntactically relevant parts of speech (POS) tag classes. These classes do not

https://doi.org/10.21862/diss-09-013-silb-etal

necessarily correspond to Hebrew orthography. In the morphology and orthography of Hebrew, words are often formed by concatenating smaller *parts*, which function as free morpho-syntactic units, each of which with its own POS tag. The basic parts are: a potentially polymorphemic stem; a template/pattern morpheme; and possibly an affix. The other morphemes within Hebrew orthographic words are clitics for certain prepositions, conjunctions, definiteness marking, and other POS (Bar-Haim, Sima'An & Winter, 2008). For the purpose of this study, we defined Hebrew clitics, such as conjunctions, articles, prepositions and so on, as words, following Silber-Varod (2013a).

The phoneme inventory of Hebrew contains five vowels and 26 consonants. There are no short–long phonemic pairs. Hebrew is a 'stressed-timed' language where primary stress can be found on one of the right-most three syllables in the word, with final stress being the most common pattern. The syllable nucleus in Hebrew is always a vowel. Word length is 2.3 syllables on average (SD 0.8) in spontaneous speech (Silber-Varod & Levy, 2014), and words rarely consist of 6 or more syllables.

The goal of this study is to analyze Hebrew PRs to see to what degree that morphology and syllable structure might influence the distribution of prolonged segments in spontaneous speech of the language. We had three main hypotheses: (i) PRs will occur mostly on the final segments; (ii) PRs will occur mostly on monosyllabic words; (iii) Prolonged vowels will be more frequent than prolonged consonants.

Method and material

Thirty-six speakers participated in this study. Ages ranged between 21 and 54 years; the median age was 30 years. 21 of the speakers were females. The session began with a structured interview comprising six fixed questions (detailed in Silber-Varod et al., 2016). The interviewer was an MA student in clinical psychology. All participants were native speakers of Hebrew living in Israel. There were no indications of hearing, 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 Sennheiser MKE 2 microphone digitized with an Icicle 48V external sound card connected to a computer. The microphone was at a fixed distance from the speaker's mouth, and the recording was carried out with a sampling frequency of 48 kHz, 16 bit sample resolution. The duration of the analyzed spontaneous narratives was about 94 minutes (ca. 3 minutes/speaker).

Target segments

Prolongations were segmented manually by a phonetician (one of the authors), a native Hebrew speaker, using Praat software (Boersma & Weenink, 2015). Although relying mostly on perception, a minimum threshold of 230 milliseconds was set, based on Silber-Varod (2013a). Vowel boundaries were marked between the onset and offset of the second formants of the vowels. Duration measurements were carried out automatically using Praat's textgrids.

The annotation of prolonged syllables resulted in 347 prolonged segments. The minimal duration was found to be 247 ms and the maximal 1.964 s. The perceived prolonged segments were then compared to a durational model of Hebrew segments in fluent speech (Modan, 2018: 124, 169–173) and to syllable durations as a function of prosodic environments (Silber-Varod, 2013a: 74–77). The comparison to the fluent durational model ruled out two segments of two different speakers because the ratios between the fluent durational model and the actual prolonged segments were above 50%. Thus, prolonged segments in the current study are *at least* twice as long as their fluent counterparts. The comparison to syllable durations as a function of prosodic environments (Silber-Varod, 2013a) ruled out four different vowels, produced by four different speakers, of which two were those ruled out also by Modan's (2018) model, and the other two fluent/prolonged ratios were 36% and 37%.

On the next step, the remaining 343 prolongations were categorized according to five parameters:

(i) Number of syllables of the word containing the prolonged segment (from 1 to 4);
(ii) Position of the target segment in the word (initial, medial, final, and monosyllabic words that consist of a single vowel);
(iii) Type of segment (vowel vs. consonant);
(iv) Duration of the prolonged segment;
(v) Sex.

Results

On average, speakers uttered 3.57 PRs per minute (ranging from 0.51 to 7.65 PRs per minute), and in total an average of 9.53 PRs per speaker (variability ranges from 1 to 30). Figure 1 presents these results. The ratio of 343 PRs is 3.5% of all word tokens.

Number of syllables in words

Table 1 shows that the number of syllables in the affected words played a role in segment prolongation. 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.

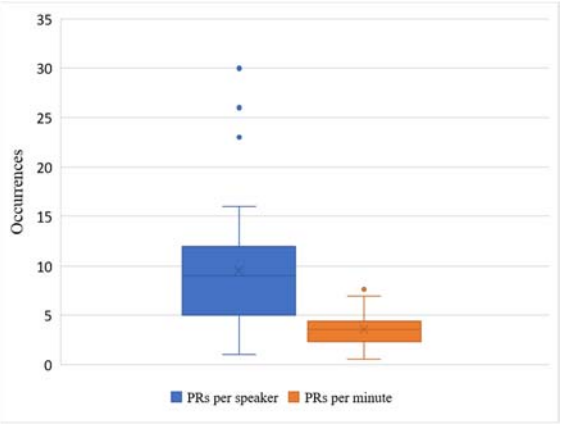


Figure 1: Prolongation distribution per speaker and per minute. Median and mean are indicated.

Table 1. Occurrences and relative frequency of prolongation as a function of number of syllables in the affected word.

Number of syllables in word	Occurrences of the words	Relative frequency (%)	CCV	CV	CVC	V	VC
1	254	74.05	5	240	1	4	4
2	51	14.87	0	43	6	1	1
3	37	10.79	0	33	2	2	0
4*	1	0.29	0	1	0	0	0
Total	**343**		5	317	9	7	5

* The 4-syllable word is a loan word.

A comparison of the relative frequencies of the syllable per prolonged word in the current study to the ratios that were found in spontaneous Hebrew (Silber-Varod and Levy, 2014), shows higher rates of prolonged monosyllables (74.05% in the current study versus 48% in spontaneous speech), and lower rates of the prolonged words with >1 syllables. As to the syllable structure of the prolonged syllables, we found five syllabic structures: CCV, CV, CVC, V, VC. Table 1 presents the distribution of each syllable structure within the number of syllables in the word.

Position

Regarding the distribution of initial-medial-final segment prolongation, our results show that in Hebrew, 97.95% of PRs are final, while the other positions are negligible (Table 2). Note that 41 cases of monosyllable CV words with [h] as the onset were categorized as Final PRs. These cases include the definite article /ha/ 'the', and the two singular third-person pronouns /hu/ 'he' and /hi/ 'she'.

Regarding lexical stress, prolongations occur mostly (74.05%) on monosyllabic function words that have no assigned stress. Within the 51 disyllabic words, we found 91.48% PRs on a stressed syllable (for example [aníː] 'I'). Within the 37 tri-syllable words, we found 75.67% PRs on an unstressed syllable (for example, [anáxnuː] 'we'). In total, PRs occur mostly on unstressed syllables (84.83%).

Table 2. PR distribution in words (occurrences)

Number of syllables in word	Initial	Medial	Final	Single V
1	1	0	249	4
2	1	1	49	0
3	0	0	37	0
4	0	0	1	0
Total	**2**	**1**	**336**	**4**

Segments

The prolonged segments are presented in Table 3. All speakers prolonged vowels (Vs), while only 13 speakers prolonged consonants (Cs). Interestingly, the most elongated segment is the vowel [e], which is also the sole realization of filled pauses in Hebrew. In the Cs section, it is not surprising to find the two nasals [m] and [n], the liquid [l] and two unvoiced fricatives [ʃ] and [x]. It is interesting to note that an extremely prolonged [xː] has only recently entered the Hebrew lexicon as an interjection which means *LOL* (Laugh Out Loud). As is seen, prolongations are distributed in a very restricted manner, as compared to the previously studied languages (see Introduction).

Table 3. Distribution of segments subject to prolongation and relative frequency given as percentages.

Vowels (*N*=324; 94.46%)	Occurrences	Relative frequency
e	166	48.40%
a	62	18.08%
i	40	11.66%
u	35	10.20%
o	21	6.12%
Consonants (*N*=19; 5.53%)		
m	12	3.50%
l	4	1.17%
ʃ	1	0.29%
n	1	0.29%
x	1	0.29%

Duration of the prolonged segments

In Figure 2, we show the results of our durational analysis, broken down for Vs and Cs. As in the previous calculation, here also we first averaged prolonged Vs and prolonged Cs per speaker, and then across speakers. The average duration of the prolonged Vs is 0.57 seconds and the average duration for prolonged Cs is 0.51 seconds. A two-tailed *t*-test showed no significant differences between the durations of Vs and Cs ($p = 0.72$).

Sex

In total, men produced 52% of the prolongations (179 versus 163 that women did), with a standard deviation of 6.69 for men and 5.89 for women. On average, men produced 11.93 PRs per interview, while women only 7.45 PRs. A one-tailed *t*-test showed that this difference was statistically

significant ($p = 0.03$). Regarding duration, on average, PR segments produced by men are 0.599 seconds and PR segments produced by women are 0.540. In an unpaired *t*-test, this difference was not found to be statistically significant ($p = 0.19$).

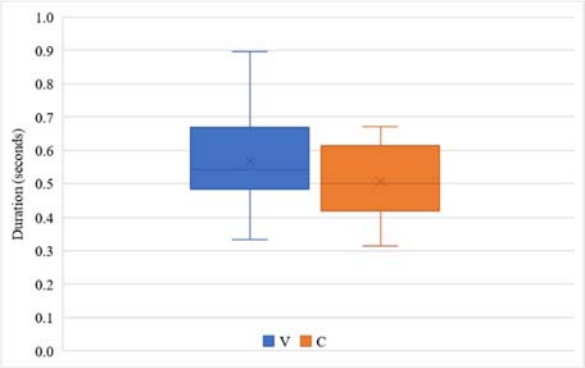


Figure 2. Durations of PRs broken down for vowels and consonants.

Discussion and conclusions

In this paper, an analysis of 343 PRs in spontaneous Hebrew resulted in several generalizations: Regarding *Distribution*, Hebrew prolongations occur 97.95% on the final segment and scarcely on initial or medial segments. Interestingly, this makes Hebrew more similar to Tok Pisin, Japanese and Mandarin Chinese than to English, German or Swedish, with a roughly 1–1–98% distribution. As for *Segments*, vowels are, on the whole and perhaps not surprisingly, more frequently prolonged than consonants in our data, unlike American English and Swedish. Regarding *stress*, PRs occur mostly on unstressed syllables. As for *Duration*, we did not find any effect of segment type (vowel or consonant) on the durations of PRs. As for *Sex*, there is a significant tendency for male speakers to produce more PRs than female speakers do. Assuming that PRs are a strategy to keep the floor, these findings support the hypothesis that men are less prone to yielding the floor in dialog (see Eklund & Wirén, 2010: 23).

Acknowledgements

We are grateful for Hamutal Kreiner from Ruppin Academic College for funding the transcriptions.

References

Bar-Haim, R., K. Sima'An & Y. Winter. 2008. Part-of-speech tagging of Modern Hebrew text. *Natural Language Engineering* 14(2): 223–251. https://doi.org/10.1017/S135132490700455X

Betz, S., R. Eklund & P. Wagner. 2017. Prolongation in German. In: R. Eklund & R. Rose (eds.), *Proceedings of DiSS 2017, the 8th Workshop on Disfluency in Spontaneous Speech*, 18–19 August 2017, KTH Royal Institute of Technology, Stockholm Sweden, *TMH-QPSR* Volume 58(1): 13–16.

Boersma, P. & D. Weenink. 2015. Praat: Doing phonetics by computer (ver. 6.0.41) http://www.praat.org/ (Accessed 25 August 2018).

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*, 87–90.

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

Eklund, R. 2004. *Disfluency in Swedish human–human and human–machine travel booking dialogues*. Ph.D. dissertation, Linköping University, Sweden.

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, vol. 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*, Lund, Sweden, 23–28.

Gósy, M. & R. Eklund. 2017. Segment prolongation in Hungarian. In: R. Eklund & R. Rose (eds.), *Proceedings of DiSS 2017, the 8th Workshop on Disfluency in Spontaneous Speech*, KTH Royal Institute of Technology, Stockholm Sweden, *TMH-QPSR* Volume 58(1): 29–32.

Gósy, M. & R. Eklund. 2018. Language-specific patterns of segment prolongation in Hungarian. *The Phonetician*, 115, 36–52.

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

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

Modan, D. 2018. *Models for the prediction of durations in Hebrew Speech*. Ph.D. dissertation, The Hebrew University of Jerusalem.

Silber-Varod, V. 2013a. The SpeeCHain Perspective: Form and Function of Prosodic Boundary Tones in Spontaneous Spoken Hebrew. LAMBERT Academic Publishing. https://doi.org/10.1075/la.215.10sil

Silber-Varod, V. 2013b. Structural analysis of prosodic pattern: The case of excessive prolongations in Israeli Hebrew. *Revista Leitura, Special Issue on Speech Prosody* 52: 271–291. https://doi.org/10.28998/2317-9945.2013v2n52p271-291

Silber-Varod, V., H. Kreiner, R. Lovett, Y. Levi-Belz & N. Amir. 2016. Do social anxiety individuals hesitate more? The prosodic profile of hesitation disfluencies in Social Anxiety Disorder individuals. *Proceedings of Speech Prosody 2016*. 31 May–3 June 2016, Boston, USA, 1211–1215. https://doi.org/10.21437/SpeechProsody.2016-249

Silber-Varod, V. & T. Levy. 2014. Intonation Unit Size in Spontaneous Hebrew: Gender and Channel Differences. *Proceedings of Speech Prosody 7*, 20–23 May 2014, Dublin, Ireland, 658–662.

Acoustic-phonetic characteristics of Thai filled pauses in monologues

Thanaporn Anansiripinyo [1,2], Chutamanee Onsuwan [1,2]
[1]Department of English and Linguistics, Faculty of Liberal Arts, Thammasat University, Bangkok, Thailand
[2]Center of Excellence in Intelligent Informatics, Speech and Language Technology, and Service Innovation, Thammasat University, Bangkok, Thailand

Abstract

Filled pause (FP) is one type of disfluent phenomena that is commonly found in everyday speech. It has been widely studied in many languages, but little is known about this topic in Thai. This work explored three important acoustic-phonetic characteristics of Thai filled pauses in monologues. To elicit target monosyllabic tokens of FPs and those of regular word (RW) counterparts, 31 Thai adult females were asked to watch two short cooking videos and describe the contents. They were also asked to read out loud target word lists. Three acoustic measures: syllable duration, first (F1) and second formant (F2) frequencies were taken from 613 tokens. Across vowel contexts, only F2, not F1, in FPs, was significantly different from that in RWs. Differences in syllable duration between RWs versus FPs were near significant. The findings suggest that Thai speakers produced FPs in a presumably different way from RWs. In: FPs, the syllable was relatively lengthened and the tongue position was moved towards the center of vowel space. Future directions include a detailed analysis of FPs in terms of amplitude, fundamental frequency, pause duration before/after fillers and other non-linguistic factors.

Introduction

It is a well-known fact that words in a language may have several meanings and some may serve more than one function. To illustrate, English words such as 'like' and 'well', which are content words/ regular words (RWs), are also used as filled pauses in some spoken contexts (e.g. casual speech). As FPs, their meanings differ greatly from the original ones. Since FPs can be observed quite often in spoken data, they could present some challenges and obstacles for speech technology that relies on Automatic Speech Recognition (ASR) system, such as automatic translator (Gabrea & O'Shaughnessy, 2000; Ogata et al., 2009; Medeiros et al., 2013). We believe that understanding underlying acoustic differences between FPs and regular words is a useful step for differentiating between them.

Even though FPs have been studied widely in many languages (e.g. O'Shaughnessy, 1992; Watanabe et al., 2004; Vasilescu, Adda-Decker & Nemoto, 2007; Gósy et al., 2014 etc.), research on FPs in Thai is insufficiently studied. Most of the studies focused on FPs used by Thai learners of English (Pletikosa & Rungrojsuwan, 2018; Williams & Korko, 2019). Only two studies (Chaimanee, 1996; Panichkul, 2003) have looked at the use of FPs in the Thai language.

Chaimanee (1996) compared frequency of occurrences of FPs in Thai conversations carried out between native (Thai) and non-native speakers (Chinese, Japanese, American, French, Frisian, and German). FPs that were found in this study include common FPs (e.g. [(ʔ)ɯ:m], and [(ʔ)ɤ̀:]) and Thai lexical FPs (e.g. [kɔ̂:] 'also', and [tɕʰâj] 'yes'). Some speakers combined Thai lexical FPs with common FPs, for example, [kʰàʔ.ʔà:] (Thai 'final particle' [kʰàʔ] and common FP [ʔà:]).

Almost a decade later, Panichkul (2003) investigated acoustic characteristics of FPs in Thai monologues focusing on their duration, amplitude, and fundamental frequency. Data were collected from 30 Thai speakers talking about a given topic; 1,162 FPs were analyzed. She found that average duration of FPs was significantly longer than that of surrounding words, while the amplitude and fundamental frequency of FPs were significantly lower. It is noteworthy that similar finding for the duration differences were found in languages such as Japanese, English, French, and Spanish (Quimbo, Kawahara & Doshita, 1998; Vasilescu, Adda-Decker & Nemoto, 2007).

However, Panichkul (2003) did not include any Thai lexical FPs in her analysis and did not analyze other important features such as formant frequencies. Vocalic hesitations have demonstrated interesting features and have been studied in many languages (e.g. Vasilescu, Adda-Decker & Nemoto, 2007). In French, it is likely to be /œ/, while in American English, it is between /ʌ/ and /æ/, and /e/ in Spanish. Comparisons of F1 and F2 between FPs and RWs also suggested that there is a small tendency for higher F1 and lower F2 in FPs in Portuguese. (Proença et al., 2013)

This work presents findings from three main characteristics of common and Thai lexical FPs which are duration and first and second formant frequencies. More importantly, our data were collected and

https://doi.org/10.21862/diss-09-014-anan-onsu

analyzed in such a way that we could make systematic comparisons between FPs and RWs. The study is part of a larger project investigating acoustic characteristics of FPs (duration, f0, F1, and F2) and other non-linguistic factors, such as speaking style, topic, and age group.

Method

Participants

Thirty-one Thai adult females (24–68 years old; mean = 42.42 ± 15.64 SD) were recruited to perform two experimental tasks: describing cooking methods and reading target word lists. Participants of the same gender were selected for convenience in the acoustic data comparison.

Speech materials

In this study, participants were asked to perform two tasks: describing cooking methods and reading target word lists. In the first task, participants were instructed to watch two short cooking videos (3–5 minutes each) with Thai audio descriptions. The two videos differ in terms of complexity level ('easy' and 'more complicated') of cooking steps and variety of ingredients. The first 'easy' video is for "Coriander smoothie" and the second 'more complicated' video is a recipe for "Chor Muang" (type of Thai sweet; violet flower-shaped steam dumpling). They all started with the 'easy' video. After finish watching each video (one time each), they were asked to repeat the cooking methods, names of ingredients, and other important information as much as they could provide.

During the second task, they were asked to read target and non-target words in a sentence frame. Target words consisted of 24 monosyllables, 14 of which are FPs (7 common FPs [(ʔ)ɯːm, (ʔ)ùːm, (ʔ)ûːm, (ʔ)ɤː, (ʔ)ɤ̀ː, (ʔ)àː, (ʔ)ɔ̀ː] and 7 Thai lexical FPs [bæ̀ːp, kɔ̂ː, kʰàʔ, kʰráp, kʰūː, lǽːw, mǔan] as shown in Table 1. Ten non-target words share similar vowels with the FPs. These items randomly appeared three times (in three different word lists) so that each participant read 72 tokens in total.

Speaking rate was not directly controlled in both tasks (analysis of speaking rate is conducted separately and not included here); participants were encouraged to talk comfortably at their normal speaking rate.

Token selection

It should be noted that FPs are produced spontaneously and quite uncontrollably in speech. For the purpose of this study, acoustic analysis and comparison were conducted only on matching monosyllables from each speaker that were produced both as FPs (in describing task) and in RWs (in reading task). As a result, only eight monosyllables satisfied the condition; 256 tokens of RWs and 357 tokens of FPs as shown in Table 2. The last column shows a number of speakers (out of 31 participants) who produced these items during both tasks.

Table 1. Target monosyllables (common and Thai lexical FPs) in word lists.

[(ʔ)ɯːm]	[bæ̀ːp] 'form, pattern'
[(ʔ)ùːm]	[kɔ̂ː] 'also'
[(ʔ)ûːm]	[kʰàʔ] 'final particle'
[(ʔ)ɤː]	[kʰráp] 'final particle'
[(ʔ)ɤ̀ː]	[kʰūː] 'is'
[(ʔ)àː]	[lǽːw] 'and'
[(ʔ)ɔ̀ː]	[mǔan] 'identical'

Table 2. Number of tokens for RWs and FPs used in acoustic analysis.

Monosyllable	No. of tokens		No. of speaker
	Regular word	Filled pause	
(ʔ)ɤː	17	8	4
(ʔ)ɤ̀ː	44	120	18
(ʔ)àː	41	91	15
(ʔ)ɔ̀ː	8	9	6
bæ̀ːp	24	33	8
kɔ̂ː	60	63	20
kʰàʔ	33	23	12
kʰūː	21	10	7
Total	256	357	

Mispronounced tokens in the reading task were excluded resulting in unequal numbers of repetitions. In terms of occurrences of FPs, the highest occurrence was for common FPs [(ʔ)ɤ̀ː], followed by [(ʔ)àː], and Thai lexical FP [kɔ̂ː] 'also'. (Frequency of occurrences of all FPs from the first task is analyzed but not included here).

Acoustic analysis: For syllable duration measurement, we measured (with Praat; Boersma & Weenink, 2017) the entire syllable unit by marking the starting point at the start of initial sound and the ending point at the last sound of vowel or final (if exists). For formant frequencies, F1 and F2 values were extracted from the mid-point of vocalic portions; the point where formant frequency is likely to be the most stable.

Statistical analysis: Syllable duration, F1 and F2 values of each token were analyzed with statistical test and average and Standard Deviation (*SD*) for each factor were calculated. N-Way Analysis of Variance (N-way ANOVA) was used to test the differences of three factors (syllable duration, F1, and F2) between the tokens occurred in RWs and FPs.

Acoustic-phonetic characteristics of Thai filled pauses

Results of each factor will be presented as follows.

Syllable duration

Figure 1 shows that in 7 out of 8 items, syllable duration of FPs were longer than that of RWs, except in [bæ̀:p]. In Table 3, average duration across all FPs was longer (339.90 ± 159.12 SD) than that of RWs (296.99 ± 85.73 SD). In addition, duration range of FPs was wider (53.50–1197.45 ms) than RWs (166.55 – 645.55 ms). However, the duration differences were near significant [$F(1,166) = 3.32$, $p = 0.0704$] (the dfs reflect number of target monosyllables that satisfied the condition, See Table 2). Post-hoc analysis indicated the same trend.

Formant frequencies (F1 and F2)

In Table 3, slightly higher average F1 and higher average F2 were found across all monosyllables (across different vowel contexts) in RWs than in FPs. N-Way Analysis showed that across vowel contexts, significant difference of F2 in FPs versus RWs was found [$F(1,166) = 5.57$, $p = 0.0194$], but not of F1 [$F(1,166) = 0.09$, $p = 0.7705$].

Interesting pattern emerges as we compare each individual word/vowel in Figure 2. In FPs (triangle shapes) (as opposed to RWs), their F2 values were located closer to the center of the vowel space. In fact, for the front vowel in [bæ̀:p], F2 in FP was lower while for the back vowels in [(ʔ)ɔ̀:] and [kɔ̂:], F2 in FP was higher (see Discussion section below).

Post-hoc analysis revealed that F2 of [bæ̀:p] in RW was significantly higher than in FP ($p < 0.01$). For F1, post-hoc analysis showed that significant differences were found in common FPs, [ʔɤ̀:] and [ʔà:]. F1 of [ʔɤ̀:] in FP was significantly higher than that in RW ($p < 0.01$) and F1 of [ʔà:] in FP was significantly lower than RW ($p < 0.05$).

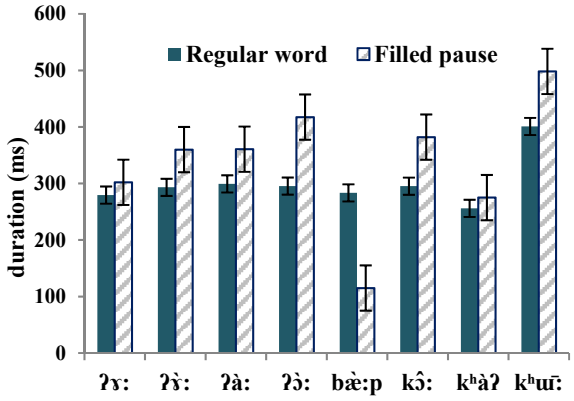


Figure 1. Average syllable duration of 8 monosyllables produced as RWs (left bar) vs. FPs (right bar). The whiskers indicate SDs

Table 3. Means and SDs of syllable duration, F1, and F2 values from 256 tokens of RWs and 482 tokens of FPs.

		Syllable duration (ms)	F1 (Hz)	F2 (Hz)
RWs	Mean	296.99	715.32	1606.79
	SD	85.73	230.00	355.06
FPs	Mean	339.90	712.82	1514.88
	SD	159.12	169.90	243.90

Discussions and future directions

We are aware that our data set was limited as naturally produced FPs are not easy to control and that the RW counterparts were produced in a connected sentence frame rather than spontaneous speech. Nevertheless, the findings revealed an interesting trend that Thai speakers produced RWs and FPs in a relatively different way. Syllables in FPs were lengthened to

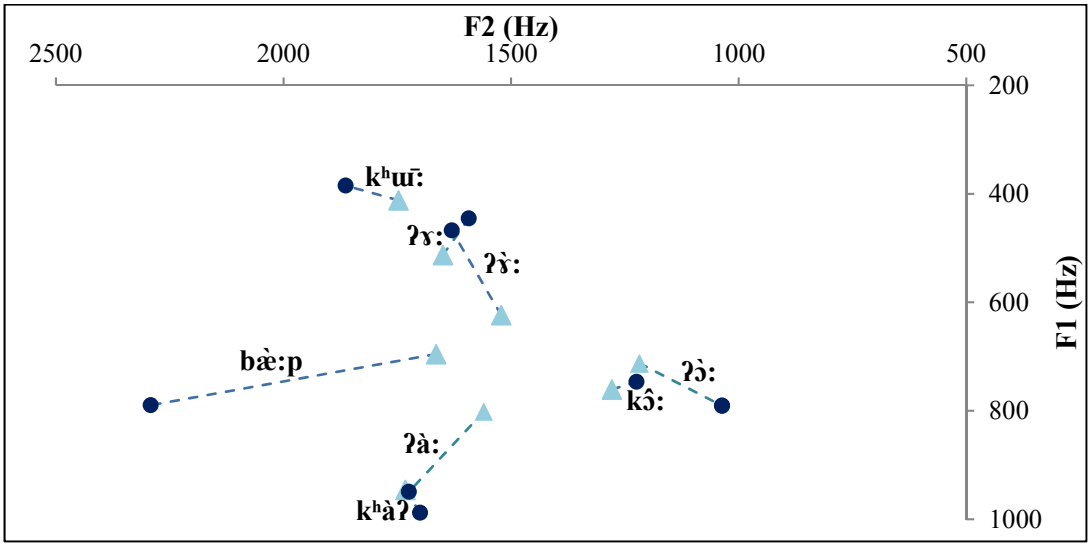


Figure 2. Average F1 and F2 values of 8 monosyllables produced in RWs and FPs. Formant frequencies were taken from the mid-point of vocalic portions

some extent and the tongue position was moved closer towards the center of vowel space. F2 dimension was the one that showed the most significant effect.

Syllable duration differences were near significant, but the results largely agreed with previous studies (Quimbo, Kawahara & Doshita, 1998; Panichkul, 2003; Vasilescu, Adda-Decker & Nemoto, 2007) that duration of FPs tends to be longer than RWs. In the case of [bæ̀:p] in our data, the only exception for the pattern, we speculated that in FPs (but not in RWs), the tokens were likely pronounced with a different vowel, a contrastive short [bæ̀p] rather than the long [bæ̀:p]. To a native ear, changing to [bæ̀p] did not change the word meaning, but made it sound less formal.

As for formant frequencies, with most studies focusing on a few vowels in FPs, their results seemed to show that FPs had higher F1 and lower F2 (Vasilescu, Adda-Decker & Nemoto, 2007; Proença et al., 2013). In our data, we examined a number of different vowels and clearer patterns appeared to emerge (see Figure 2). In varying degrees, each vowel was produced differently in RWs versus FPs, but in FPs, the tongue position apparently (not surprisingly) moved towards the center; e.g. the vowel in [bæ̀:p] was pronounced more like [ə] or [a] in FPs.

Future directions include a detailed analysis of FPs in terms of amplitude, fundamental frequency, pause duration before/after fillers and other non-linguistic factors (such as age and gender). Preliminary results of FP frequency of occurrences seem to show that level of complexity of the task/topic has a noticeable effect. As the task becomes more complicated and challenging, speakers are more likely to show various disfluency strategies such as silent pauses, lengthening and FPs.

Acknowledgment

The project was partially supported by Center of Excellence in Intelligent Informatics, Speech and Language Technology, and Service Innovation, Thammasat University.

References

Boersma, P. & D. Weenink. 2017. Praat: Doing phonetics by computer (ver. 6.0.56) http://www.praat.org/ (Accessed 22 July 2017).

Chaimanee, N. 1996. Communicative pauses in Thai. In: *Pan-Asiatic linguistics: Proceedings of the 4th International Symposium on Language and Linguistics*, Mahidol University, 8–19 January 1996, Bangkok, Thailand, 174–182.

Gabrea, M. & D. O'Shaughnessy. 2000. Detection of filled pauses in spontaneous conversational speech. In: B. Yuan, T. Huang, & X. Tang (eds.), *Proceedings of the International Congress of Phonetic Sciences*, 16–20 October 2000, Beijing, China, 678–681.

Goto, M., K. Itou & S. Hayamizu. 1999. A real-time filled pause detection system for spontaneous speech recognition. In: G. Olaszy, G. Németh, & K. Erdöhegyi (eds.), *Proceedings of the 6th European Conference on Speech Communication and Technology*, 5–9 September 1999, Budapest, Hungary, 227–230.

Gósy, M., J. Bóna, A. Beke & V. Horváth. 2014. Phonetic characteristics of filled pauses: The effects of speakers' age. In: S. Fuchs, M. Grice, A. Hermes, L. Lancia, & D. Mücke (eds.), *Proceedings of the 10th International Seminar on Speech Production (ISSP)*, 5–8 May 2014, Cologne, Germany, 150–153.

Medeiros, H., H. Moniz, F. Batista, I. Trancoso & H. Meinedo. 2013. Experiments on automatic detection of filled pauses using prosodic features. In: F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, & P. Perrier (eds.), *Proceedings of the International Speech Communication Association*, 25–29 August 2013, Lyon, France, 2629–2633.

Ogata, J., M. Goto & K. Itou. 2009. The use of acoustically detected filled and silent pauses in spontaneous speech recognition. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 19-24 April 2009, Taipei, Taiwan, 4305-4308. https://doi.org/10.1109/ICASSP.2009.4960581

O'Shaughnessy, D. 1992. Recognition of hesitations in spontaneous speech. In: *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 23–26 March 1992, San Francisco, USA, 521–524. https://doi.org/10.1109/ICASSP.1992.225857

Panichkul, S. 2003. *An acoustic study of Thai pause fillers in relation to their syntactic positions in monologues*. Master's dissertation, Chulalongkorn University, Bangkok, Thailand.

Pletikosa, J. & S. Rungrojsuwan. 2018. The use of filled pauses in monologue by Thai learners of English with different English language proficiency levels. *Vacana Journal of Language and Linguistics*, 6(1): 20–35.

Proença, J., D. Celorico, A. Veiga, S. Candeias & F. Perdigão. 2013. Acoustical characterization of vocalic fillers in European Portuguese. In: R. Eklund (ed.), *Proceedings of DiSS 2013, the 6th Workshop on Disfluency in Spontaneous Speech* and *TMH-QPSR Volume 54(1)*, KTH Royal Institute of Technology, 21–23 August 2013, Stockholm, Sweden, 54(1): 63–66.

Quimbo, F. C., T. Kawahara & S. Doshita. 1998. Prosodic analysis of fillers and self-repair in Japanese speech. In: R. H. Mannell & J. Robert-Ribes (eds.), *Proceedings of the International Speech Communication Association*, 30 November – 4 December 1998, Sydney, Australia, 3313–3316.

Vasilescu, I., M. Adda-Decker & R. Nemeto. 2007. Acoustic and prosodic characteristics of vocalic hesitations across languages. *Scientific Report 2007*.

Watanabe, M., Y. Den, K. Hirose & N. Minematsu. 2004. Clause types and filed pauses in Japanese spontaneous monologues. In: S. H. Kim, & D. H. Youn (eds.), *Proceedings of the International Speech Communication Association*, 4–8 October 2004, Jeju, Korea, 2981–2984.

Williams, S. A. & M. Korko. 2019. Pause behavior within reformulations and the proficiency level of second language learners of English. *Applied Psycholinguistics*, 40: 723–742. https://doi.org/10.1017/S0142716418000802

Disfluencies in spontaneous speech in easy and adverse communicative situations: The effect of age

Linda Taschenberger, Outi Tuomainen and Valerie Hazan
Department of Speech Hearing and Phonetic Sciences, UCL, London, UK

Abstract

Disfluencies are a pervasive feature of speech communication. Their function in communication is still widely discussed with some proposing that their usage might aid understanding. Accordingly, talkers may produce more disfluencies when conversing in adverse communicative situations, e.g. in background noise. Moreover, increasing age may have an effect on disfluency use as older adults report particular difficulties when communicating in adverse conditions. In this study, we elicited spontaneous speech via a problem-solving task from four different age groups (19–76 years old) to investigate the effect of energetic and informational maskers on the use of filled pauses (FPs), and its interaction with age. Measures of disfluency rates, effort ratings, and communication efficiency were obtained. Results show that, against our predictions, FP usage may decrease in adverse conditions. Moreover, age does not play a great role in adults with normal hearing. The results indicate that individuals differ greatly in their disfluency adaptations, utilising different strategies to overcome challenging communicative situations.

Introduction

Spontaneous communication is marked by its speech flow interruptions (Bortfeld et al., 2001). Two kinds of pauses are frequently examined in disfluency research: silent pauses and filled pauses. Silent pauses are periods of non-articulation from a talker. Filled pauses, on the other hand, are periods of articulation of non-lexical content (Clark & Fox Tree, 2002), e.g. 'uh' and 'uhm' in English.

The effects of filled pauses (FPs) are widely discussed, with some proposing that they are used to buy the speaker time in order to plan their utterances (e.g. Tottie, 2014; Jucker, 2015) or hold the floor in conversation (Shriberg, 1994; Fox Tree & Clark, 1997). Others believe that FPs may prepare listeners for unexpected words and thereby actually aid understanding (Fox Tree, 1995; Arnold, Fagnano & Tanenhaus, 2003). Bortfeld et al. (2001) suggest that disfluencies may be affected by cognitive, social, and situational aspects. Shriberg (2001: 156) argues along similar lines stating that they "are related to the speaking environment in which they arise". Many have found an increase in disfluency (DF) use as task difficulty increases (e.g. Levin, Silverman & Ford, 1967). Some studies have also shown that talkers become more disfluent when they are speaking in background noise (Southwood & Dagenais, 2001). Furthermore, there are well-documented changes in both speech production and perception with increasing age. Some find that older talkers generate more disfluencies compared to younger adults (Bortfeld et al., 2001). Older adults (OAs) also often report having particular difficulty communicating in challenging situations (e.g. in noise). However, this interaction is not yet well explored; it is therefore not known whether OAs become more disfluent than younger adults (YAs) when communicating in adverse listening conditions.

The aim of the current study is therefore twofold: firstly, to explore the effect of filled pauses in the context of adverse listening conditions. Secondly, to investigate the role age plays in disfluency use. To investigate the impact of listening difficulty, we manipulated the type of interaction during completion of the same task: in the presence of 'energetic' (EM) and 'informational' masking (IM). It has been well established that the type of background noise can differentially affect listeners, with IM causing greater interference than EM (Rudner et al., 2012). We therefore recorded spoken interactions taking place (a) in quiet with no interference present, (b) in noise with no informational content (EM), and (c) in background speech (IM).

To elicit spontaneous interactive speech, we used a problem-solving 'spot-the-difference' picture task (diapix, van Engen et al., 2010) under these three different listening conditions. We measured rate of disfluencies, communication efficiency (i.e. time it took to find differences), and self-rated listening effort and concentration. The research questions addressed are: 1) Does the rate of disfluencies differ across noise conditions? 2) Does the rate of disfluencies differ across age? 3) Do disfluencies have an effect on the communication efficiency and perceived effort and concentration of conversation?

We predict that the rate of FPs will be higher in the adverse conditions compared to the quiet condition. Furthermore, we predict that IM will result in higher use than EM as it is known to cause more interference. Regarding age, we predict that it will have an effect on frequency of FP use with older adults using more

https://doi.org/10.21862/diss-09-015-tasc-etal

than younger adults in all conditions. Similarly to the age differences found in other speech strategies employed in noise, talkers may adopt an increased use of FPs as a strategy to overcome the effect of background interference. Concerning the perceived effort of conversation, we do not have strong hypotheses: effort could either be rated lower and communication may be more efficient if an interlocutor uses more FPs as this may prepare listeners for unexpected words and thereby aid understanding. Alternatively, effort could be rated higher and communication could be less efficient if FPs are a marker of the talker struggling with the task at hand.

Method

Participants

54 monolingual native speakers of Standard Southern English participated in the study. They were aged between 19–26 years (Younger Adults, YA, $N = 20$, 10 F, Mean age 21.8 years), 30–49 years (Middle Aged, MA, $N = 12$, 8 F, Mean age 42.4 years), 50–64 (Older Middle Aged, OMA, $N = 10$, 10 F, Mean age 61.8 years) and 65–76 years (Older Adults, OA, $N = 12$, 10 F, Mean age 71.3 years). Participants were tested in pairs of the same sex within the same age band. All participants had normal hearing thresholds (< 25 dB HL) across the 0.25–4 kHz range and reported no history of speech and language impairments or neurological trauma

Procedure

During the audio recordings, participants sat in separate acoustically-shielded rooms and communicated via headsets fitted with a cardioid microphone (Beyerdynamic DT297) whilst playing interactive Diapix games (Baker & Hazan, 2011) on a desktop PC. Participants were given different versions of the same picture scenes (e.g. Figure 1) and told that they had 10 minutes to find the 12 differences between the pictures.

Each talker was recorded on a separate channel at a 44,100 Hz (16 bit) sampling rate using a Fireface audio interface and Audacity audio software. One of the talkers (designated *Talker A*) was told to lead the interactions. The other (*Talker B*) was a more passive participant who mainly responded to queries by Talker A. All participants carried out both talker roles. After completion of every Diapix task, participants were asked to rate their effort and concentration (11-point Likert scale: "Did you have to put in a lot of effort to understand your partner?", 0 = lots of effort, 10 = no effort; "Did you have to concentrate very hard to understand your partner?", 0 = concentrate hard, 10 = not concentrate).

The picture task was carried out in four listening conditions (three of which are reported here) affecting

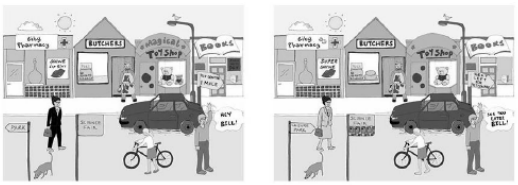


Figure 1. A DiapixUK picture pair.

both participants: both speakers in i) quiet, ii) EM with no informational content ('speech-shaped' noise), and iii) IM that is semantically related to the picture description task (i.e. talking about the same picture). The IM condition was a 3-talker masker consisting of a male, a female and a child speaker. The picture and noise condition orders were randomised. The apparatus and procedure were identical to Hazan, Tuomainen and Taschenberger (2019). The data were collected as part of a wider protocol, but only relevant tasks are described here.

Data processing

All recordings were automatically transcribed using a speech recognition system by Speechmatics and then manually corrected in Praat (Boersma, 2001). Annotation of filled pauses was performed against the word level transcriptions by the first and second author using the spellings UH, UHM, UM, ER, ERM. From the recordings we calculated measures that reflect i) Talker A's amount of FPs as a percentage of total utterances spoken (as recording durations were not consistent), ii) communication efficiency (i.e. time in seconds from start to finding 8th difference), and iii) listening effort and concentration ratings for Talker B.

Results

Filled pauses

Overall, filled pause usage reflected other reported findings with FPs taking up an average of 3.86% of all utterances (Clark, 1994). To establish whether FP use differed across noise conditions, a repeated measures ANOVA was carried out for the within-subject factor listening Condition (3: QUIET, EM, IM). In general, use of FPs was higher in the quiet condition ($M = 4.27\%$, $SD = 2.20$) than in both of the adverse conditions (EM: $M = 3.67\%$, $SD = 1.90$; IM: $M = 3.66\%$, $SD = 1.93$). This main effect of Condition was significant ($F(3, 159) = 3.929$, $p < 0.01$) against our hypothesis that disfluency use would be greater in adverse conditions. Post hoc analyses showed no difference across adverse conditions, but that this significance was driven by the differences to QUIET.

Age analyses were based on linear mixed-effects modelling using the `lme` function in the `nlme` package for `R` (Version 1.1.463). The best-fitting model for each individual analysis was chosen with hierarchical

approaches, that is, adding one predictor at a time to a baseline model that includes no predictors other than the intercept. Condition (3 levels) and Age (continuous; centred at the mean across age bands) were entered one by one as fixed effects and Participant as random effect. Likelihood ratio tests were used to determine which effects were needed in the model. Age showed a statistically significant interaction effect ($p < 0.01$) in only the IM condition (see Figure 2). Here, FP usage declined with increasing age.

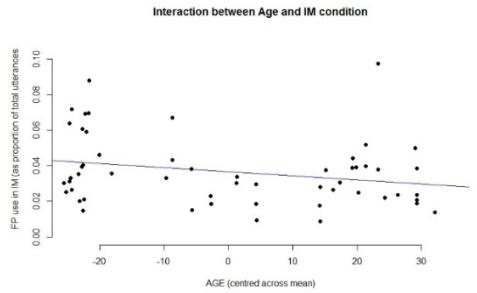


Figure 2. Interaction effect between Age and IM condition for FP use (as proportion of total utterances).

As these findings go against our predictions and all show high variance, we decided to investigate the individual differences in the data. We firstly calculated the percentage change relative to FP use in QUIET for all participants. This showed that participants greatly differed in their use of FPs overall and across conditions (see Figure 3). Some increased their usage in adverse conditions, while others decreased it. We then calculated individual *z*-scores for each participant for each condition, with a *z*-score at ±1 SD taken as a meaningful difference, denoting that the individual was making marked changes in their FP production across conditions (i.e. outside the 15th–85th percentile range). This showed that in EM, 12 individuals differed from the population mean, five were below the 15th percentile and seven above the 85th percentile. This indicated a change of −50% to −100% and 39% to 113% change relative to QUIET, respectively. In IM, 15 individuals differed from the population mean, with seven found under the 15th percentile and eight over the 85th percentile. This indicated a change of −50% to −70% and 42% to 100%, respectively. It is also of interest to see if

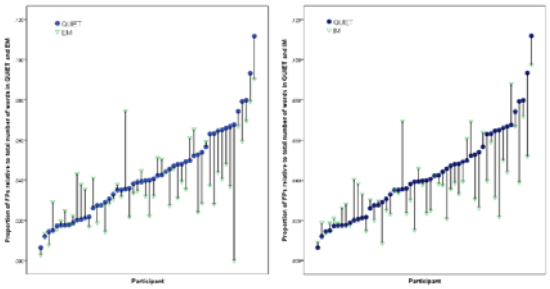


Figure 3. FP rate for each participant between QUIET and EM (left) and QUIET and IM (right).

speakers were consistent in their FP strategy across adverse conditions. Five participants meaningfully increased their FP use in both adverse conditions. One individual meaningfully decreased FP rate in both adverse conditions. One individual decreased in EM and increased in IM. The rest of the participants only showed a significant change in FP.

Perceived effort and communication efficiency

In order to investigate whether ratings given immediately after completing Diapix were associated with FP use, we carried out bivariate Pearson's correlation calculations for mean ratings as a function of partner's FP use. These analyses were done for two ratings (described in "Data processing" above) evaluating both effort and concentration. For Effort ratings, the only significant correlation at the $p < 0.01$ level was for the IM condition, with higher effort ratings in higher FP use instances ($r = -0.34$). With regards to communicative efficiency, we correlated whether Talker A's FP use had an effect on the time it took to find eight differences in the pictures. This factor only played a minor role in QUIET ($p = 0.036$, $r^2 = 0.08$). None of the adverse conditions were influenced by FP use here. Overall, these correlation analyses showed only weak relationships with most going in the opposite direction of our predictions.

Discussion

It has often been found that talkers increase their disfluency use in conversational speech as task difficulty (e.g. Levin, Silverman & Ford, 1967) or background noise increases (e.g. Southwood & Dagenais, 2001) or with increasing age (e.g. Bortfeld et al., 2001). In the current study which recorded conversational speech with communicative intent in easy and adverse speaking conditions in speakers spanning the adult age range, all with normal hearing thresholds up to 4 kHz, we could not replicate these findings. Overall, use of filled pauses decreased in the more challenging situations. Additionally, age played only a minor role. However, our data set showed large variance: analyses of individual differences suggest that speakers adopt differing strategies in their speech adaptations. When communication becomes effortful, talkers need to make various adjustments to their speech production to aid listeners' understanding, for example, by speaking more slowly and reducing complexity of their utterances (Gagné et al., 1994). There are large individual differences in how effectively clear speech is produced (e.g. Hazan et al., 2018). This may extend to disfluency use, too. Background interference may not affect everyone to the same degree and result in the same production strategies. It has indeed been widely noted that

disfluency rates vary between corpora (e.g. Branigan et al., 1999; Lickley, 2001) with different overall dialogue tasks and speaker roles greatly affecting the way disfluencies are produced.

Importantly, we solely investigated the use of filled pauses. It may be other types of disfluencies that increase in more challenging situations (e.g. hesitations or repetitions). Complementary analyses on silent pauses will therefore be carried out and be presented at the conference. It may be the case that the less filled pauses are used in adverse conditions, the more silent pauses are utilised. This additional information will tell us more about the strategies adopted in adverse speaking conditions.

However, our lack of a significant age effect might also be due to the fact that a decline in sensory acuity (e.g. hearing ability) may be a contributing factor. Tuomainen and Hazan (2017) found a significant relationship between hearing thresholds and disfluency usage with poorer hearing resulting in more disfluent conversations. As our participant group all had hearing thresholds under 25 dB, this might have eliminated the conflating age effect found in some other studies.

Acknowledgements

This work was supported by the Economic and Social Research Council [grant number ES/P002803/1].

References

Arnold, J. E., M. Fagnano & M. K. Tanenhaus. 2003. Disfluencies signal theee, um, new information. *Journal of* Psycholinguistic *Research* 32(1): 25–36. https://doi.org/10.1023/A:1021980931292

Baker, R. & V. Hazan. 2011. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods* 43(3): 761–770. https://doi.org/10.3758/s13428-011-0075-y

Boersma, P. & D. Weenink. 2008. Praat: Doing phonetics by computer (version 6.0.46). http://www.praat.org/ (accessed 31 May 2019).

Bortfeld, H., S. D. Leon, J. Bloom, M. F. Schober & S. Brennan. 2001. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech* 44(Pt 2): 123–147. https://doi.org/10.1177/00238309010440020101

Branigan, H., R. Lickley & D. McKelvie. 1999. Non-linguistic influences on rates of disfluency in spontaneous speech. In: J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville & A. C. Bailey (eds.): *Proceedings of ICPhS 1999, the 14th International Congress of Phonetic Sciences*, 1–7 August 1999, San Francisco, CA, USA, volume 1, 387–390.

Clark, H. 1994. Managing problems in speaking. *Speech Communication* 15(3–4): 243–250. https://doi.org/10.1016/0167-6393(94)90075-2

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

Fox Tree, J. E. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language* 34(6): 709–738. https://doi.org/10.1006/jmla.1995.1032

Fox Tree, J. E. & H. H. Clark. 1997. Pronouncing 'the' as 'thee' to signal problems in speaking. *Cognition* 62(2): 151–167. https://doi.org/10.1016/S0010-0277(96)00781-0

Gagné, J. P., V. M. Masterson, K. G. Munhall, N. Bilida & C. Querengesser. 1994. Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech. *Journal of the* Academy *of Rehabilitative* Audiology 27: 135–158.

Hazan, V., O. Tuomainen, J. Kim, C. Davis, B. Sheffield & D. Brungart. 2018. Clear speech adaptations in spontaneous speech produced by young and older adults. *The Journal of the Acoustical Society of America*, 144(3): 1331–1346. https://doi.org/10.1121/1.5053218

Hazan, V., O. Tuomainen & L. Taschenberger. 2019. Speech communication in background noise: effects of aging. In: S. Calhoun, P. Escudero, M. Tabain & P. Warren (eds.) *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, 805–809.

Jucker, A. H. 2015. Uh and Um as Planners in the Corpus of Historical American English. In: I. Taavitsainen, M. Kytö, C. Claridge and J. Smith (eds.): *Developments in English: expanding electronic evidence*. Cambridge University Press: Cambridge, 162–167. https://doi.org/10.1017/CBO9781139833882.013

Levin, H., I. Silverman & B. L. Ford. 1967. Hesitations in children's speech during explanation and description. *Journal of Verbal Learning & Verbal Behavior* 6(4), 560–564. https://doi.org/10.1016/S0022-5371(67)80017-3

Lickley, R. J. 2001. Dialogue moves and disfluency rates. In: M. Core (ed.) *DISS'01 Disfluency in Spontaneous Speech*, 29–31 August 2001, Edinburgh, UK, 93–96.

Rudner, M., T. Lunner, T. Behrens, E. S. Thorén & J. Rönnberg. 2012. Working memory capacity may influence perceived effort during aided speech recognition in noise. *Journal of the American Academy of Audiology* 23(8): 577–589. https://doi.org/10.3766/jaaa.23.7.7

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

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

Southwood, M. H. & P. Dagenais. 2001. The role of attention in apraxic errors. *Clinical Linguistics and Phonetics,* 15(1–2)*:* 113–116. https://doi.org/10.3109/02699200109167641

Speechmatics (Cloud transcription service). https://www.speechmatics.com (accessed 31 May 2019).

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

Tuomainen, O. & V. Hazan. 2017. Disfluencies in spontaneous speech in younger and older adults in easy and difficult communicative situations. *Workshop on Challenges in Analysis and Processing of Spontaneous Speech,* 14–17 May 2017, Budapest, Hungary, 7–9.

Van Engen, K. J., M. Baese-Berk, R. E. Baker, A. Choi, M. Kim & A. R. Bradlow. 2010. The Wildcat Corpus of Native- and Foreign-Accented English: communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech,* 53(Pt 4): 510–540. https://doi.org/10.1177/0023830910372495

Vowel lengthening — Effect of position, age, and phonological quantity

Valéria Krepsz
Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary

Abstract

The present research examined the effect of phrase-final lengthening on the spectral structure of vowels in the spontaneous speech of children and adults. Three Hungarian vowel pairs (in quantity pairs) were analyzed in two positions: in the middle of the phrase and at the end of the phrase. The effect of lengthening on the spectral structure of the vowels were already be detected in four-year-olds. However, its extent was strongly correlated with the articulation aspects of the vowels. There was a discrepancy in the tendencies of the lengthening's effect between the two groups of children and the adults, presumably due to different linguistic experience, inaccuracy of articulation, and significant individual differences.

Introduction

Examination of prosodic features of speech is greatly encumbered by the fact that these elements are multi-dimensional, and also contain several acoustic parameters that overlap in terms of function and appearance (Cutler et al., 1997). In addition to the meaning-modifying and meaning-carrier functions of these prosodic elements, they provide a dividing function in the utterances. Although these factors are present not only at the end of utterances, earlier research findings suggest that combinations of these individual elements occur more frequently with a boundary marker function. Previous studies mostly agree on the list of prosodic elements that can function as boundary markers, however, it is also important to emphasize that realization can vary depending on a number of factors (Wightman et al., 1992). Some of the main boundary markers are (silent) pauses, changes in fundamental frequency, changes of voice quality, and phrase-final lengthening.

The current research focuses on the phenomenon of phrase-final lengthening (PhFL) in the spontaneous speech of Hungarian monolingual children and adults. The results of previous studies suggested that phrase-final lengthening can be observed in English-speaking children by the age of 2, and confirmed that PhFL as a learnt prosodic feature has a close connection with the quality and amount of linguistic experience (Snow, 1994). Other studies suggested that the occurrence of PhFL was still not systematic in English-speaking children at the age of 8, and the individual differences were substantial (Dankovičová et al., 2004). Krepsz et al. (in press) have confirmed the appearance of vowel elongation in the last syllable in pre-pausal position already at the age of 3. The results showed a significant difference in the duration of the 4 most common Hungarian vowels depending on the quality and position (initial, medial and final occurrence in the phrase) of the vowels. The duration of vowels were gradually decreasing with age, independently from the other examined features.

The articulation of vowels at the segmental level is basically described with three features: the horizontal and vertical position of the tongue (for the latter, some authors use the term closeness or openness) relative to its resting position, and lip activity. Considering the differences in quality and duration, we can assume that there are only 9 vowel qualities in Hungarian, which (illustrated by the long member of the phonological pairs) are: /i: y: u: e: ø: o: ɛ a: ɒ/. However, long vowels differ from their short pairs more or less in their quality, the difference being strongly dependent on the tongue's vertical position (cf. Kassai, 1998; Gósy, 2004). This phenomenon can be grasped well within the framework of the H&H theory (Lindblom, 1963) that raises the question of whether the structure of hypo- and hyperarticulation vowels are different. Therefore the question of the current research whether the target configurations of Hungarian short and long vowels are different, and if so, whether the short vowels can be described as over-configured gestures of their long pairs (for a detailed description of the problem in Hungarian, see the following study). Mády's (2008) research, based on the examination of two speakers with the use of electromagnetic articulographs, found that /o/ and /o:/ vowels differ from each other in their height (openness), there was a clear difference between the articulation of the vowels /ɛ/ and /e:/, especially in terms of their height (openness).

Based on the previous results, the aim of the present study was to investigate whether there was a difference in the formant structure of short and long vowel pairs in different positions (with lengthening in phrase-final position, without lengthening in

https://doi.org/10.21862/diss-09-016-krepsz

phrase-medial position). According to our hypotheses: i) the lengthening effect on the formant structure of the short vowels will be less pronounced than on that of the long ones, ii) the effect of lengthening will be different in the realization of mid and open vowels, iii) differences will be smaller in children than in adults.

Subjects, material and method

Participants

Data were analyzed from 20 monolingual, Hungarian speaking girls aged 4 (n = 10) and 6 (n = 10) from the GABI database (Bóna et al., 2014). The participants had, as indicated by their parents, typical cognitive skills, and no speech, language or hearing problems. The socio-economic status of the participants was not controlled; however, all of the children were recruited from public kindergartens in Hungary. The children's data were compared to sound samples of young women (n = 10; aged between 25 and 35 years), the samples having been selected from the BEA database (Gósy, 2013).

Method and material

In this paper, we investigate the spontaneous speech of subjects from all three age groups. All speakers talked about their family and free time activities; children also talked about the rules of their favorite games, while adults talked about their studies and jobs. The total length of the recorded material analyzed here was 143 minutes 34 seconds, 8 minutes 45 seconds of spontaneous speech per adult and 3 minutes 22 seconds per child (on average).

The choice of the basic unit of annotation can influence the examined values, such as phrase-final lengthening. However, speech production of four- and six-year-olds is characterized by short speech periods and long pauses, even though there are great individual differences. Therefore we considered speech units from pause to pause as the basic unit of this segmentation. In this way, the data of speakers of different ages could be compared. The annotation of speech units and vowels was carried out manually under Praat (Boersma & Weenink, 2018). The segmentation of the vowels was based on their second formants, supported by visual analysis of their respective wide-band spectrograms and waveforms.

To control for factors affecting vowel duration, the following criteria were used for selecting vowels for analysis: a) Only allophones of /ɒ/, /a:/, /ɛ/, /e:/, /o/, and /o:/ were analyzed to control for vowel quality, b) Vowels were chosen from two positions within the utterances: the absolute end of the phrase (before the break, phrase-final position) and from the middle of the words (phrase-medial position), c) Vowels were chosen only from closed syllables, and d) from unstressed positions.

The data set included a total of 4094 vowels, for which duration as well as first and second formants were analyzed. In addition, the following formula was used to examine the effect of elongation of vowels at phrase-final position: the duration of the speech unit was divided by the number of syllables (until the examined syllable, which was excluded), which gave the articulation rate in syllable per second. Then the duration of the last syllable was divided by the above mentioned articulation rate. This number showed the ratio of the (phrase-final) lengthening (compared to the phrase-medial position).

For statistical analyses, we used the R program (linear mixed model; Venables, 2018). The independent variables were 'age' with three levels (4- and 6-year-olds, and adult), 'position' with two levels (phrase-final, PhF and phrase-medial, PhM) and 'vowel quality' with six levels, 'vowel quantity' with two levels (short and long vowels). The person of the speaker was the random factor. The dependent variables were duration, the F1 and the F2 value of the vowels and 'Euclidean distance'.

Results

The phenomenon of PhFL was detectable irrespective of the age of the speakers and the quality of the vowels. In the case of short vowels, the ratio of lengthening was higher (0.012 on average), and it was lower (0.008 on average) in the case of long vowels (Figures 1 and 2). There was a gradual decrease with the increase in the speaker's age: the greatest lengthening was found in four-year-olds (0.013 on average), then in six-year-olds (0.011 on average) and the smallest in the group of adult speakers (0.009 on average).

The statistical analysis showed significant difference in the duration of the vowels depending on the position [$F(1,4094) = 26.142$; $p = 0.001$] and age [$F(1,4094) = 14.826$; $p = 0.019$]: the duration of vowels shortened with age, however, the pairwise

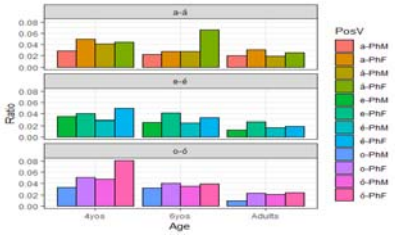


Figure 1. The extent of lengthening (ratios) depending on the speaker's age and the quality of the vowels.

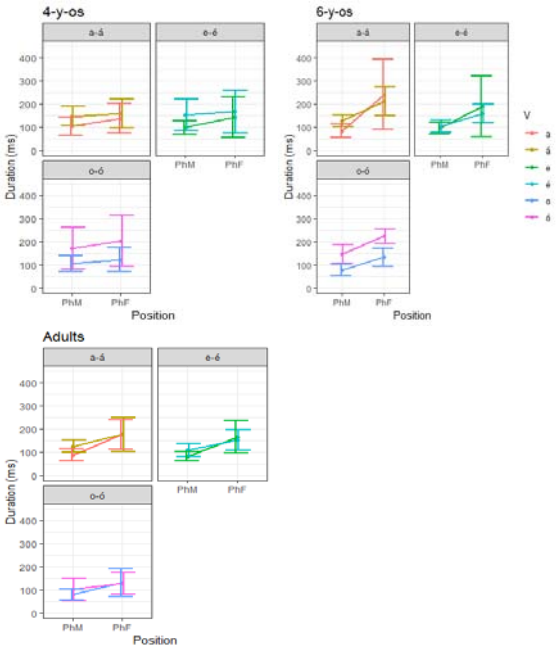


Figure 2. The duration of vowels by the age of the speaker and the quality and quantity of the vowels.

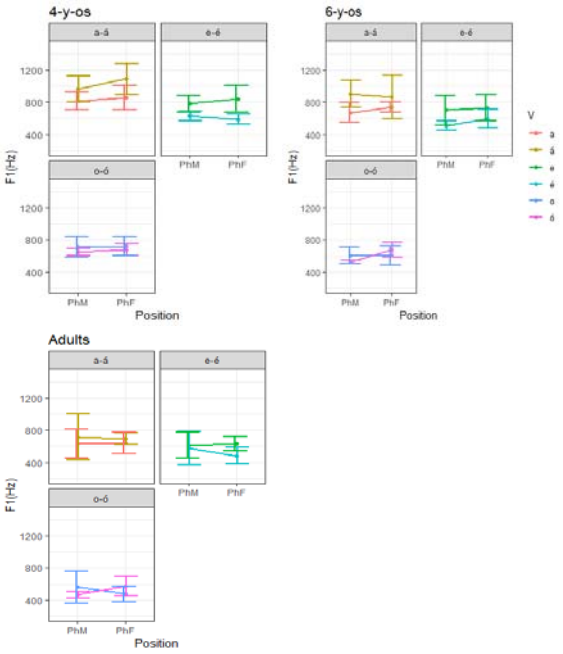


Figure 3. F1 values depending on the age of the speakers and the quality and quantity of the vowels.

comparison proved difference only between children and adults.

The quality of the vowel and the interaction of the factors did not prove to be decisive in vowel lengthening.

Formant values were analyzed depending on position, vowel quality and the speaker's age. According to the statistical analysis, there was significant difference in F1 values by age [$F(2, 4094) = 10.546$; $p = 0.022$], and vowel quality [$F(5, 4094) = 33.317$; $p = 0.014$]. Although position did not prove to be a decisive factor in itself, tendentious differences regarding vowel quality were

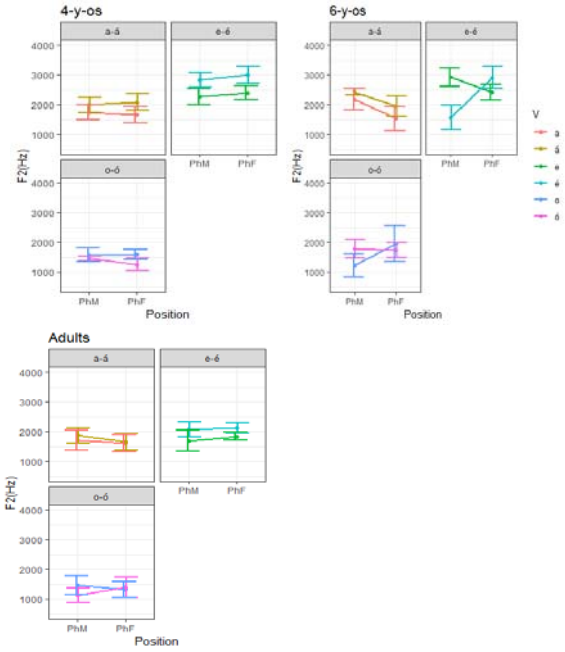


Figure 4. F2 values depending on the age of the speakers and the quality and quantity of the vowels.

observed in the development of F1 values. As Figure 3 shows, the tendencies of /ɒ/ and /ɛ/ vowels were clear: in the case of phrase-final position, the value of F1 was higher.

The realization of /a:/ and /e:/ vowels was determined by the age of the speakers: for the vowel /a:/, difference between the phrase-final and phrase-medial positions was found only in four-year-olds, where the lengthened realization had higher F1 values. For the vowel /e:/, higher F1 values were found at the phrase-medial position in the group of 4-year-olds and adults; and at the phrase-final position in the group of 6-year-olds. There was only a slight difference between the two positions for /o/; while for /o:/, although the differences were small, lower mean F1 values were measured in the phrase-final position in all three age groups.

F2 values showed no significant difference by position either: values of F2 showed diversity by the quality and quantity of vowels and by age. Similarly to the first formant, there was a significant difference in the second formant by the quality of the vowels [$F(5, 4094) = 44.434$; $p = 0.007$] and the speaker's age [$F(2, 4094) = 9.842$; $p = 0.030$], and the interaction of the position and vowel quality [$F(5, 4094) = 4.438$; $p = 0.034$], as well as the triple interaction of position, vowel quality and age were also significant [$F(21, 4094) = 9.931$; $p = 0.015$].

In the case of /ɒ/ and /a:/ the tendency was clear: the average of F2 decreased at the phrase-final position in all three age groups (Figure 4). The difference according to position was minimal for /ɛ/ and /e:/, the average difference was only 15 Hz. In the case of /o o:/ pairs the trend was different

depending on age: in four-year-olds and adults we measured higher F2 values at the phrase-final position, while the tendency was the opposite in six-year-olds. In the case of /oː/, there was a decreasing tendency for children and an increasing tendency for adults.

The Euclidean distance from the center of the vowel area was not influenced by the position.

Conclusions

The present study analyzed the effect of the position (PhF or PhM) on the duration and formant structure (F1 and F2) of 3 Hungarian quantity vowel pairs in the spontaneous speech of children and adults. The phenomenon of PhFL was detectable independently from the other analyzed factors and the rate of the lengthening was higher in children than in adults. Although there was no significant difference in either F1 or F2 values according to position, (limited) conclusions about the changes in vowel quality can be drawn. In addition, it is important to emphasize that from the change of vowel formant structure, we can only draw conclusions about the articulation of vowels indirectly and with constraints. The lengthening was most determinative in the case of /ɒ/ and /aː/, it is likely that the pronunciation of both vowels became more open and the position of the tongue shifted backwards. In addition, the change of the first formant of the /ɛ/ vowel may indicate that it has also been more openly pronounced. Although the quantity of the vowel as a factor was also not determinative for either of the two formant values, the change in the formant values of the short vowels was less variable with regard to the speakers' age. The variance of the formants was greater in the case of higher heights, while in the case of lower heights, greater stability and lesser age sensitivity were observed. Age proved to be a determinative factor for most of the analyzed features. There was a discrepancy in the tendencies of the lengthening between the two groups of children, presumably due to different linguistic experience, inaccuracy of articulation, and significant individual differences. In addition, it is important to emphasize that the imitation of tongue movements is difficult as they mostly take place inside the mouth cavity. Therefore, the acoustic differences resulting from tongue movement appeared and become similar to adults' at a late station of the language acquisition.

Acknowledgements

This research was supported by the Hungarian National Research, Development and Innovation Office of Hungary, project No. NKFIH-K-120234.

References

Boersma, P. & D. Weenink. 2018. Praat: Doing phonetics by computer (version 6.0.37). http://www.praat.org/ (accessed 14 November 2018)

Bóna, J., A. Imre, A. Markó, V. Váradi & M. Gósy. 2014. GABI – Gyermeknyelvi beszédAdatBázis és Információtár [GABI - Children's Speech and Informational Database]. *Beszédkutatás* 22: 246–251.

Cutler, A., D. Dahan & W. Van Donselaar. 1997. Prosody in the comprehension of spoken language: A literature review. *Language and speech* 40(2): 141–201. https://doi.org/10.1177/002383099704000203

Dankovičová, J., K. Pigott, B. Wells & S. Peppé. 2004. Temporal markers of prosodic boundaries in children's speech production. *Journal of the International Phonetic Associaion* 34(1), 17–36. https://doi.org/10.1017/S0025100304001525

Gósy, M. 2004. *Fonetika, a beszéd tudománya*. Budapest: Osiris.

Gósy, M. 2013. BEA – A multifunctional Hungarian spoken language database. *Phonetician* n. 105–106: 50–61.

Kassai, I. 1998. *Fonetika*. Budapest: Nemzeti Tankönyvkiadó.

Krepsz, V., V. Horváth, M. Gósy & A. Huszár. In press. Magánhangzók temporális mintázata az anyanyelv-elsajátításban [Temporal pattern of vowels during first language acquisition].

Lindblom, B. 1963. Spectrographic Study of Vowel Reduction. *The Journal of the Acoustical Society of America* 35(11): 1773–1781. https://doi.org/10.1121/1.1918816

Mády, K. 2008. Magyar magánhangzók vizsgálata elektromágneses artikulográffal normál és gyors beszédben [Analysis of Hungarian vowels with electromagnetic articulograph in normal and fast speech]. *Beszédkutatás* 16: 52–66.

Snow, D. 1994. Phrase-Final Syllable Lengthening and Intonation in Early Child Speech. *Journal of Speech Language and Hearing Research* 37(4): 831–840. https://doi.org/10.1044/jshr.3704.831

Venables, W. N. & D. M. Smith. R Core Team. 2018. *An Introduction to R — Notes on R: a programming environment for data analysis and graphics* (version 3.5.0). https://cran.r-project.org/ (accessed 4 October 2018).

Wightman, C. W., S. Shattuck-Hufnagel, M. Ostendorf & P. J. Price. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* vol. 91(3): 1707–1717. https://doi.org/10.1121/1.402450

Temporal characteristics of teenagers' spontaneous speech and topic based narratives produced during school lessons

Mária Laczkó
Faculty of Pedagogy, Kaposvár University, Kaposvár, Hungary

Abstract

The aim of this presentation is to analyse the articulation and speech rates of teenagers and the types of pauses in their spontaneous speech and topic based narratives during school lessons. The speech samples were analysed in terms of temporal characteristics by Praat program. The results showed the different tempo values and various function of filled pauses in the examnined situations.

Introduction

The tempo of speech is the rate at which utterances and their smaller units are pronounced. It is defined as speaking and articulation rates. Speaking rate refers to entire speaking phase including pauses versus articulation rate which refers to phases of articulation excluding pauses (Fletcher, 2010; Jacewicz, Fox & Wei, 2009). So the speech tempo indicates all of the number of different speech units pronounced by the speaker with the pauses versus the articulation rate indicates the speed of the number of speech units produced in the time actually taken to articulate it excluding the pauses from it (Gósy, 2004: 203).

Both the speaking and articulation rate are influenced by internal and external factors like age, gender, individuality and the topic of the speech, the speaking context, type of the text or the speech style (Bóna, 2010; Gocsál, 2000; de Andrade & de Oliveira Martins, 2007; Jacewicz, Fox & Wei, 2010; Laczkó, 2019; Menyhárt, 2000; Olaszy, 2006; Oyer-Deal, 1985; Quené, 2008; Torre–Barlow, 2009; Váradi–Beke, 2013).

On the basis of these findings the tempo can be changed paralel with the age and the teenagers' tempo categories seem to be fastest among the different aged people.

The scholars also emphasized the rising tempo at the beginning of teenage years both in international and Hungarian examinations (de Andrade & de Oliveira Martins, 2007; Laczkó, 1991; Menyhárt, 2000; Neuberger, 2014). The Hungarian students' speed of their spontaneous speech were also different in terms of the age (Laczkó, 2009).

The actual research question is what the speech tempo of teenagers in the communication situation like which require different cognitive strategies and activities than spontaneous speech and how it is characterized by various types of the pauses.

This presentation is focused on the temporal analysis of topic based narratives produced by the students during school lessons. This kind of situation is different from spontaneous speech regarding the speech planning, access of lexemas and articulation processes. In the situation of the topic based narrative (as the common responses of the students in teaching–learning process), planning, conceptualisation is not simultaneous with the formulation and articulation. During the planning process the topic is exactly defined as it is based on the actual teaching material. The linguistic form is also defined because of the required parameters of narratives (the order of events, the time of them and the interrelations (Bruner, 1994; Neisser, 1994)), but the articulation is done in the given moment.

Our hypotheses were as follows. H1) The planning process of topic based narratives is more complex for the students than spontaneous speech. H2) It can be followed in different temporal parameters (tempo categories, the type of the pauses, function of filled pauses) of topic based narratives and spontaneous speech.

Method and material

In order to discuss the hypothesis the series of experiment was carried out with the participation of teenagers The average age of them was 15.4 and 17.2 year. The students are studying in the secondary school, they all had normal hearing and intelligence with typical language development.

For the examination spontaneous speech samples and topic based narratives produced during the lessons were digitally recorded. In spontaneous speech the students had to speak about their free time activities, versus topic based narratives which were oral responses (story telling) based on the actual themes learnt by the students during the previous lessons and they had to speak about them in terms of given aspect

The time which was given to students to speak was approximately 3 minutes per person in both of the examined situations.

For the analysis the speech rates (the total number of sounds divided by total speaking time with the pauses), articulation rates (the total number of

https://doi.org/10.21862/diss-09-017-laczko

sounds divided by total speaking time without pauses), the ratio of unfilled and filled pauses were calculated in both types of speeches. The duration of different types of pauses and the function of filled pauses (speaking intention, error and repair, uncertainty (Levelt, 1989; Horváth, 2010)) was also analysed in each situations among the speakers. For the acoustic analysis the Praat program (Boersma, 2001) was used, the statistical analysis was done by the SPSS 13.00 version.

The tempo categories were measured by the number of sounds per seconds, the duration of pauses was given in milliseconds.

However the number of the students was only 5-5 in both communication situations in the different age groups, the same students took part in them. The all number of students was 20.

The time of topic based narratives among the 15 years was 8 minutes 27 seconds, and it was 8 minutes 42.7 seconds among the 17 years. The time of spontaneous speech of 15 years was 10 minutes 46.6 seconds versus the 17 years where it was 10 minutes 44.8 seconds. The average time of a speaker was roughly the same, almost 2 minutes long in topic based narratives and a little bit longer in spontaneous speech independently the age.

Results

The tempo data analysis

There was similar tendency in both age groups. The speech rates and articulaton rates were much lower in topic based narratives than in spontaneous speech (Figure 1).

The differences between the speech rates is almost 1.5 sound/sec, versus the articulation rates, where it is almost 2 sound/sec. The differences were proved by the statistical analysis (paired-samples *t*-test: $t(3)=-9.107$, $p=0.003$).

The individual tempo categories also showed the differences between the two kind of speeches. Among topic based narratives produced by 15 years the slowest speech rate was 5.4 sound/sec, the fastest was 8.91 sound/sec (*SD*: 0.911). These tempo categories among the 17 years were 3.88 sound/sec, and 6.78 sound/sec (*SD*:1.122). In topic based narratives of 15 years the lowest articulation rate was 8.53 sound/sec, the fastest one was 12.16 sound/sec (*SD*: 1.035). The articulation rates among the 17 years were 5.48 sound/sec and 10.94 sound/sec (*SD*: 1.623). In spontaneous speech of 15 years the lowest speech tempo was 7.18 sound/sec, the fastest was 9.49 sound/sec (*SD*: 0.720). These rates among the 17 years were 5.71 sound/sec and 9.02 sound/sec (*SD*: 1.02). In spontaneous speech of 15 years the lowest and the fastest articulation rates were 10.68 sound/sec and 12.88 sound/sec (*SD*: 0.832), among the 17 years these were 7.3 sound/sec and 13.25 sound/sec (*SD*: 1.477). So the individual tempo values could also prove the slower speed of topic based narratives comparing them to spontaneous speech, and the tendency could occur independently the age.

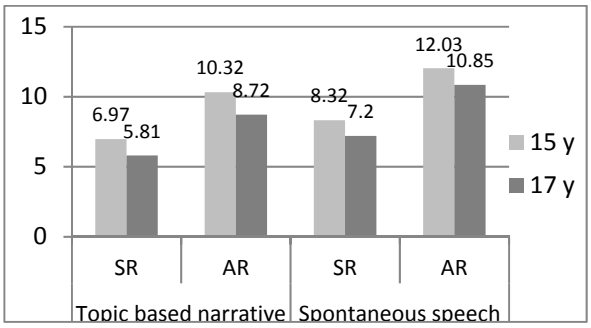


Figure 1. The articulation rate (AR) and speech rate (SR in the two kind of speeches (sound/sec)

The ratio of types of pauses was similar in the two age groups in the examined situations. In topic based narratives of 15 years the ratio of unfilled pauses was 90.7%, and 92.1% among the 17 years. The ratio of filled pauses was 9.3% in the younger group, and 7.9% in the elder one. In spontaneous speech the ratio of unfilled pauses was a little bit higher, 95.5% among the 15 years, and 96.7% among the 17 years. The ratio of unfilled pauses were a little bit lower than in topic based narratives, 4.5% in the younger group and 3.3% in the elder one.

The data of the duration of pauses showed the same tendency in both age groups. In topic based narratives the unfilled and filled pauses were much longer than in spontaneous speech (Figure 2). The differences was also proved by statistical analysis ((Paired-Samples *t*-test: $t(3)=4.205$, $p=0.025$).

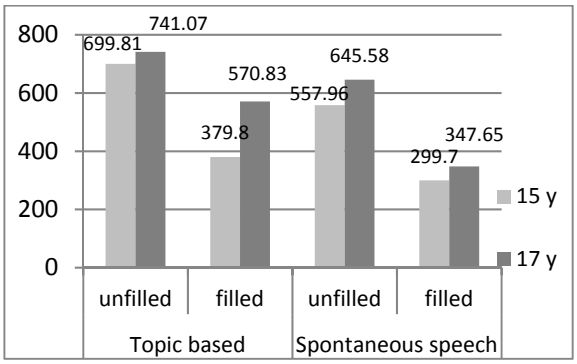


Figure 2. The average duration of the pauses (msec).

In topic based narratives of 15 years the duration of unfilled pauses were 140 ms longer than in their spontaneous speech. It was 100 ms longer among the 17 years. In terms of the filled pauses the differences can be followed mainly among the 17 years as their

filled pauses were almost 130 ms longer than in their spontaneous speech. Among the 15 years' narratives the filled pauses were 80 ms longer than in their spontaneous speech.

The analysis of filled pauses/forms and functions

The most frequent realisation form of filled pauses (Figure 3) is *œ* in both of the two types of speeches and in both of the two age groups.

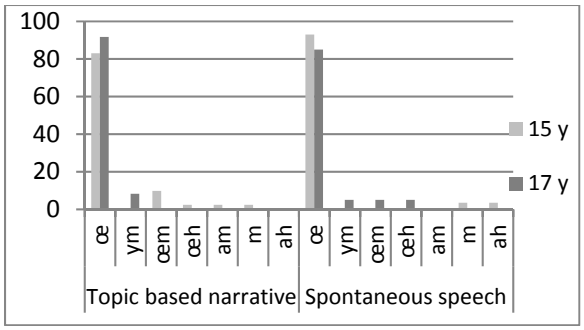


Figure 3. The realisation of filled pauses (%)

The other kind of realisation is describing or the 15 years or the 17 years. Among the 17 years the types of *ym* is the second frequent category mainly in topic based narratives. The types of *œm* has the second place in terms of the frequency. With this ratio it is occuring only in 15 years' topic based narratives. It also describes the 17 years' spontaneous speech but the ratio is a little bit lower. In the elder group there is again one category (*œh)*, but it describes only their spontaneous speech. The distribution of other kind of filled pauses is really low and it describes only the 15 years' topic based narratives and/or their spontaneous speech.

The analysis of the function of filled pauses (Figure 4) showed the opposite tendency in the two kind of speeches.

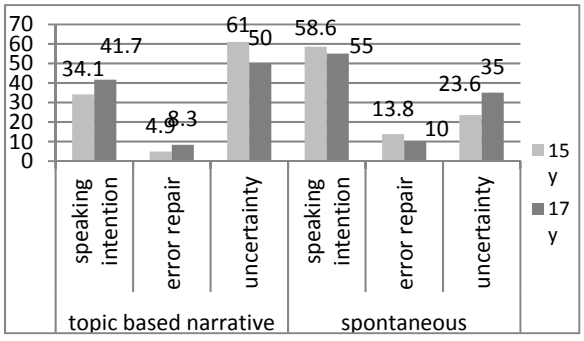


Figure 4. The function of filled pauses (%)

In topic based narratives the students used the types of filled pauses in the function of uncertainty in the highest ratio independently the age versus spontaneous speech where the highest ratios were in the function of speaking intention in both age groups. The types of filled pauses were used for the error repair mainly in spontaneous speech.

The duration of types of filled pauses in the examined three functions was also analysed (Figure 5).

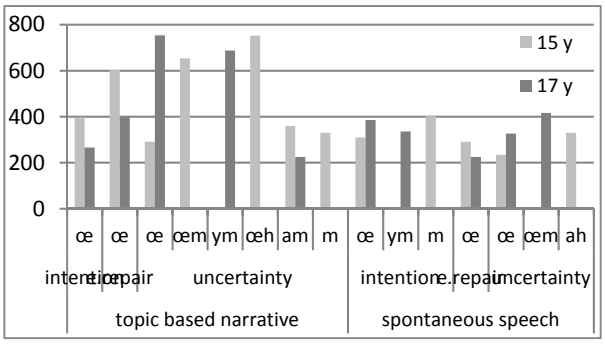


Figure 5. The duration of types of filled pauses in different function

In the function of speaking intention the students used only *œ* in topic based narratives versus spontaneous speech where other types were also used. The duration of them was longer in the 15 years' topic based narratives and shorter in the 17 years' spontaneous speech. In the function of error repair the students used only the type *œ*. The duration of them was longer in topic based narratives mainly among the 15 years. For the uncertainty the wide range of filled pauses was used and duration of them was also longer in topic based narratives.

Discussion and conclusion

The aim of this research was to prove that topic based narrative is more complex task for the secondary school students to produce than spontaneous speech and it can be reflected in their temporal characteristics.

The data showed topic based narratives' much slower speed than it was in spontaneous speech. Their slower speech and articulation rates are characterized by significantly longer duration of both unfilled and filled pauses than in spontaneous speech. The difference between the two kind of speeches was also proved in terms of types and function of filled pauses (see H2). The most frequent forms of them was almost the same in the two kinds of speech like in our previous findings (Laczkó, 2009), but the distribution of their function was different. In topic based narratives the leader position of filled pauses is the function of uncertainty versus spontaneous speech where the speaking intention. In the different function the various types of them had also different durations.

The data can support a lot of difficulties of teenagers to produce topic based narratives as the oral responses from the teaching material (see H1). Consequently the data obtained can predict the learning difficulties in the subjects which need oral expressions as their slow tempo might have the interrelation with their small vocabularies and their lexical acces difficulties. However it must be controlled with more students, we emphasize the more need of oral presentations in teaching-learning process during the lessons and the need of reading.

References

Boersma, P. & D. Weenink. 2008. Praat: Doing phonetics by computer (version 5.1.41). http://www.praat.org/ (accessed 8 December 2010).

Bóna, J. 2010. Beszédtervezési folyamatok az életkor és a beszédstílus függvényében. *Magyar Nyelvőr*. 134: 332–341.

Bruner, J. 1994. The "remembered" self. In: U. Neisser & R. Fivush (eds.): *The remembering self: Construction and accuracy in the self-narrative.* New York: Cambridge University Press, 41–54. https://doi.org/10.1017/CBO9780511752858.005

Fletcher, J. 2010 (2nd edition). The prosody of speech: Timing and rhytm. In: W. J. Hardcastle, J. Laver & F. E. Gibbon, (eds.): *The Handbook of Phonetic Sciences*, 521–602. Oxford: Wiley-Blackwell. https://doi.org/10.1002/9781444317251.ch15

Furquin de Andrade, C. & Vanessza de Oliveira, M. 2007. Fluency variation in adolescents. *Clinical Linguistics and Phonetics* 21: 771–782. https://doi.org/10.1080/02699200701502161

Gocsál, Á. 2000. A beszéd időviszonyai különböző életkorú személyeknél [Speech timing in individuals of different ages]. *Beszédkutatás* 8: 39–50.

Gósy M. 2004. *Fonetika, a beszéd tudománya*. Budapest: Osiris Kiadó.

Horváth V. 2010. Filled pauses in Hungarian: their phonetic form and functions. *Acta Linguistics Hungarica* 57(2–3): 288–306. https://doi.org/10.1556/ALing.57.2010.2-3.6

Jacewicz, E., R. A. Fox, & L. Wei. 2010. Between speaker and within-speaker variation in speech tempo of American English. *Journal of the Acoustical Society of America* 128: 839–850. https://doi.org/10.1121/1.3459842

.

Jacewicz, E., R. A. Fox, C. O'Neill & J. Salmon. 2009. Articulation rate across dialect, age and gender. *Language Variation and Change* 21: 233–256. https://doi.org/10.1017/S0954394509990093

Laczkó, M. 1991. Interrelation of articulation rate and pauses in children's speech. *Temporal Factors in speech,* Budapest: MTA, Nyelvtudományi Intézete.139–151.

Laczkó, M. 2009. Középiskolai tanulók spontán beszédének temporális jellegzetességei [Temporal characteristics of high school students' spontaneous speech]. *Magyar Nyelvőr*133: 447–67.

Laczkó, M. 2019. The Temporal Characteristics of Teenagers in Spontaneous and Rhetorical Speeches. *Journal of Linguistics and Literature* 3(1): 29–34. https://doi.org/10.12691/jll-3-1-5

Levelt, W. J. M. (1989). *Speaking. From intention to articulation*. Cambridge: MA: A Bradford Book.

Menyhárt, K. 2000. A beszéd temporális sajátosságai kétnyelvűeknél (kisiskoláskortól időskorig) [Temporal characteristics of speech in bilinguals (from pre-school age to old age)]. *Beszédkutatás* 8: 51–62.

Neisser, U. 1994. Self-narratives: True and false. In: U. Neisser & R. Fivush (eds.): *The remembering self: Construction and accuracy in the self-narrative*, 1–18. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511752858.003

Neuberger, T. 2014. *A spontán beszéd sajátosságai gyermekkorban* [Characteristics of spontaneous speech in childhood]. Budapest: Eötvös Kiadó.

Olaszy, G. 2006. Prozódiai szerkezetek jellemzése a hírfelolvasásban, a mesemondásban, a novella és a reklámok felolvasásában [Characterization of prosodic structures in news reading, storytelling, short story and advertisement reading]. *Beszédkutatás* 14: 21–50.

Oyer, H.J. & V. Deal L. 1985. Temporal aspects of speech and the aging process. *Fólia Phoniatrica* 37: 109–112. https://doi.org/10.1159/000265788

Quené, H.. 2008. Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech. *Journal of the Acoustical Society of America* 123: 1104–1113. https://doi.org/10.1121/1.2821762

Torre III, P. & J. A. Barlow. 2009. Age-related changes in acoustic characteristics of adult speech. *Journal of Communication Disorders* 42: 324–333. https://doi.org/10.1016/j.jcomdis.2009.03.001

Váradi, V. & A. Beke. 2013. Az artikulációs tempó variabilitása felolvasásban [Variability of articulation rate in reading]. *Beszédkutatás* 23: 26–41.

Pausing and disfluencies in elderly speech: Longitudinal case studies

Borbála Keszler [1] and Judit Bóna [2]
[1]Department of Modern Hungarian Linguistics, ELTE Eötvös Loránd University, Budapest, Hungary
[2]Department of Applied Linguistics and Phonetics, ELTE Eötvös Loránd University, Budapest, Hungary

Abstract

The aim of this paper was to investigate the changes in fluency of speech during ageing. The novelty of the examination is that this is a longitudinal study: it analyses the speech of 7 speakers from middle or young-old age to old-old age. Pausing strategies and frequency of disfluencies were analyzed. Results show that active aging helps to preserve certain parameters of speech characteristics of young speakers.

Introduction

Occurrence of disfluencies in spontaneous speech is influenced by several factors like speaker's age (e.g. de Andrade & de Oliveira Martins, 2010) and speech task (e.g. Beke et al., 2014). There are several studies in the literature about the changes in elderly speech compared to the speech of young adults. We know that during ageing, speech rate and articulation rate decelerate (e.g. Duchin & Mysak, 1987), f0 and voice quality change; or articulation becomes less accurate (Torre & Barlow, 2009). These changes are due to altered cognitive, hormonal and psychological functions and aging of speech organs (Hnath-Chisolm et al., 2003).

There are relatively few studies about disfluencies of elderly speech. One of its most important characteristics is word-finding difficulty (e.g. Burke et al., 1991). Some authors found that there were no differences between the speech of young and old speakers in the frequency of disfluencies (Duchin & Mysak, 1987; Searl et al., 2002; de Andrade & de Oliveira Martins, 2010). Other authors found that old speakers produced more disfluencies than young speakers did (Yairi & Clifton, 1972). Analyzing the speech of seven mentally intact 100–103-year-old speakers, it was found that disfluencies occured with the same frequency in their speech as in the speech of 70–80–90-year-old speakers (Searl, Gabel & Fulks, 2002). Similar results were found in the study of de Andrade and de Oliveira Martins (2010), in their research there was no significant difference between the speech of 60, 70 and 80+ year-old people, although there was an increasing tendency of the disruption rates along the decades.

However, the above-mentioned studies were cross-sectional examinations. The novelty of this study is that it analyses the occurrence of disfluencies in speech samples of the same speakers in a longitudinal examination. At the time of the first recordings the speakers were already middle-aged or young-old, while at the time of the third recordings they were old-old. Participants were researchers and/or teachers who were quite active professionally despite their old age. The questions were the followings: 1) How does the frequency of disfluencies change during ageing? 2) What kinds of disfluencies appear in different ages? 3) Does the examined group's speech show any difference from that of the average speaker discussed in the literature? (Given that this study deals with a quite specific population.)

We had two hypotheses: 1) Fluency of speech changes during ageing also in the various stages of elderly life. The older the speaker, the more frequent are the disfluencies. 2) The biggest difference will be measured at the oldest age compared to the others.

Methods

Speech samples were selected from the Spoken Language Database of the Department of Modern Hungarian Linguistics (ELTE Eötvös Loránd University). Speech recordings of 7 native Hungarian researchers and/or teachers (1 female, 6 males) were analysed. Speech samples of the same speaker were recorded three different times. At the first time speakers were middle-aged or young-old, the second recording was taken about 10 years or more later, at the age of 70 or above (except for the female speaker), while the third recording was taken at the age of 75+ (Table 1).

Recordings contain spontaneous speech samples: birthday interviews, interviews at award ceremonies, public speeches like comments at conferences and at other professional gatherings. While they are public speeches, their characteristics might differ from those of more informal speeches. However, they weren't previously planned speeches. 4–5 minutes of speech samples were analyzed at each time by each speaker.

Speech samples were annotated using Praat 5.0 software (Boersma & Weenink, 2008), i.e. speech

https://doi.org/10.21862/diss-09-018-kesz-bona

Table 1. Age of participants at the time of the three recordings (years) (F = female, M = male).

Speaker	First time	Second time	Third time
F1	54	64	76
M1	62	72	81
M2	62	72	83
M3	61	72	82
M4	61	70	76
M5	49	70	82
M6	59	77	90

units (between two pauses) and pauses were segmented and annotated. After that, the duration of speech units and pauses were printed out in an Excel spreadsheet with a script, and the number of syllables of the speech samples were automatically counted. Based on the data, the following parameters were calculated: proportion of pauses in the total speaking time, frequency of pauses (number of pauses in 100 syllables), mean duration of pauses, frequency of filled pauses and their proportion in the total pausing time, and frequency of all disfluencies (number of disfluencies in 100 syllables). Filled pauses were considered in phonetic sense, i.e. pauses that are filled with sounds (not words) (Fletcher, 2010).

The frequency of all disfluencies (calculated per 100 words) was defined for all speakers. Each occurrence and type of disfluencies were identified and coded by both authors. The rate of agreement was 98% between the two coders. The following types of disfluencies were analyzed: filled pauses, filler words, word- or phrase-repetitions, part-word repetitions, lengthenings, pause-within-the-words, revisions (Roberts et al., 2009).

Data were examined and compared by descriptive statistic methods in case of each speaker and each recording. Other statistical analysis was not carried out because the age of the speakers was different, so they couldn't form a homogeneous group. So, the data were examined as case studies.

Results

Figure 1 shows the frequency of pauses in 100 syllables. There is a slight increase of the frequency in the speech of 5 speakers compared to the first measurement, while the frequency decreased or hardly changed in the speech of two speakers. The connection between the pauses and the actual speech situation is confirmed by the fact that the value measured at the second time is quite varied: it is the lowest or it is between the values of the first and the third measurements.

The proportion of pauses in the total speaking time also depended on the speaker and speech

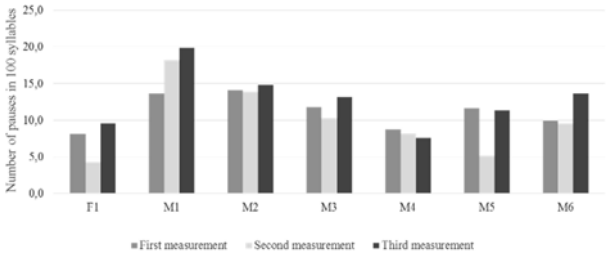


Figure 1. Number of pauses in 100 syllables.

situation (Figure 2). Values ranged from 11.7 to 26.4% at the first measurement, 12.3 to 33.5% at the second measurement, and 14.7 to 34.1% at the third measurement. This parameter showed a large increase with age by some speakers (F1, M1, M6), while it did not or only slightly changed in case of others (M2, M3, M4, M5) in the third measurement compared to the first measurement. Data from the second measurement also varied widely, indicating that age is only one factor among many others that influence temporal parameters in speech.

It should also be emphasized that the results of any speaker measured at any time (even the proportions measured in the oldest ages) are not higher than the values characteristic of native Hungarian young speakers (Bóna, 2014).

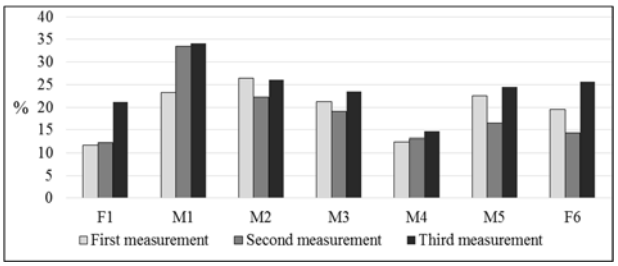


Figure 2. Proportion of pauses in the total speaking time per speaker and per measurement.

The average duration of pauses (Table 2) was longer for five speakers (F1, M1, M4, M5 and M6) at the third measurement than at the first measurement. For M2 and M3, shorter average duration of pauses was reported in the oldest age than at the first measurement.

Figure 3 shows the frequency of filled pauses. In four speakers (F1, M3, M4, M6), the change at the three measurements is very small, while in three speakers filled pauses were more frequent at the third measurement than at the first measurement. The effect of the given speech situation is indicated by the fact that in M2, filled pauses were the most frequent at the second measurement. So, it did not depend on the speaker's age.

The proportion of filled pauses in the total amount of pauses also changed mostly depending on the speaker and the actual speech situation. This

Table 2. Mean duration of pauses (F = female, M = male).

Speaker	Mean duration of pauses (ms)		
	First time	Second time	Third time
F1	427	345	538
M1	448	626	622
M2	650	624	627
M3	593	537	578
M4	329	364	367
M5	489	386	590
M6	424	316	495

proportion was very low for each speaker. In one speaker (F1), filled pauses appeared only in the speech sample recorded at the oldest age, and in a very small ratio (only 0.9% of the total pause time). In the case of four speakers (M1, M2, M3 and M5), proportion of filled pauses increased in the oldest age compared to the youngest age, but the dependence of this parameter on the speech situation is well indicated by the highest proportion at the second measurement in M2 and M5. In M4 and M6, the proportion of filled pauses was either reduced or unchanged at the third measurement compared to the first measurement.

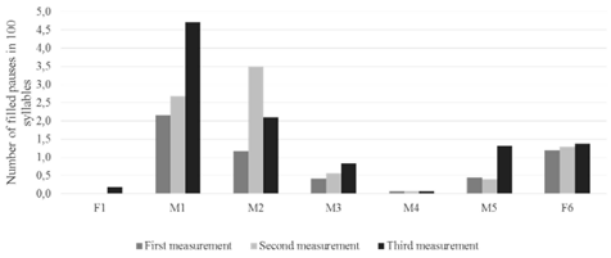


Figure 3. Number of filled pauses in 100 syllables.

In the frequency of disfluencies, we do not see any tendency, either (Figure 4). In four speakers (F1, M1, M3, M5) there was a slight but more frequent occurrence of disfluencies, while in three speakers it decreased (M6) or almost did not change (M2 and M4) at the third measurement compared to the first one. Since the data of the second measurement varied in different ways, they confirm that the frequency of disfluencies also largely depends on the current speech situation, the current state of the speaker and the topic.

Discussion and conclusion

This study analyzed the changes of fluency of speech in speech samples of 7 speakers across several decades. We had two hypotheses. One of them was partly confirmed while the other wasn't.

The first hypothesis was that the older the speaker, the more frequent the disfluencies are. This was not

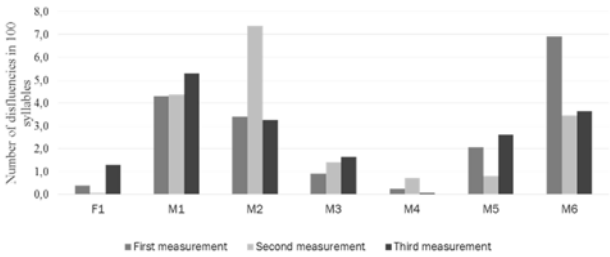


Figure 4. Number of disfluencies in 100 syllables.

confirmed. There was no clear connection between ageing and fluency. Although the speech of certain speakers became less fluent, there were other speakers who produced similarly fluent speech at the third measurement compared to the first one. Obviously, we don't know if the speakers spoke with similar fluency and pausing in every speech situation at the given age. In other words, one recording is only relevant for that very speech situation. However, this means that there were at least three speakers who were able to speak with the same fluency both at the time of the first and the last measurement. According to our expectations, the proportion of pauses would be higher in parallel with ageing. This parameter did not change in case of one speaker, changed only slightly in case of three, and changed considerably in case of another three speakers. We also hypothesized that pauses would be more frequent in older age. This is proved in case of four speakers, while three other speakers produced pauses less frequently than in their younger age. These results show that age-dependent changes show great individual variability. The most coherent changes were measured in the mean duration of pauses. This duration became longer by the oldest age in five speakers (but the increase was not linear with ageing), while in case of other two speakers it decreased slightly.

According to the second hypothesis, the greatest change was expected in the oldest age. This was only partially confirmed and mostly typical for the ratio of pauses. There were no changes in all parameters and in all speakers. Moreover, the direction of the changes was not always as it had been expected.

Results led to the following conclusion. Age does not affect the fluency of speech for every speakers. It can be assumed that the background of the data is that active life, frequent public speaking helps to preserve the mobility of speech organs and the speed of speech planning processes. A particularly important result is that the data of the very old speakers in our study were similar to those of young adults. This shows that, although the literature suggests that aging results in a slowing down of speech tempo and more pauses, some speakers can preserve a speech similar to that of young speakers.

However, some changes were found in pausing strategies (proportion of pauses and their average duration). This is due to several factors. On the one hand, during ageing (some parts of) the planning processes might become slower, and the speakers try to gain time for these processes by taking longer silent pauses. This time is enough for further planning—speakers do not need to take more pauses at older age. On the other hand, it can also be related to breathing. It was not analyzed in this study, but from an earlier study (Bóna, 2018) we know that older people breath more often (audibly) during speaking than young people. The breath-taking pauses are also statistically longer than breathless pauses (Gyarmathy, 2019).

Our results confirm the experience that active aging helps to preserve certain parameters of speech characteristics of young speakers. In addition, research shows that speakers who are accustomed to public speaking are less disfluent than someone who started public speaking later in life. Data indicate that practice patterns have a significant effect on the fluency characteristics of public speaking performance, as speakers who started practicing earlier were less disfluent than those who started later (Goberman et al., 2011), at least professional or good speakers are expected to be more fluent (Das et al., 2019). This seems to be true even in the very old age. As the analysis of the same speakers' recordings across decades is rare in both Hungarian and international literature, our results provide important new insights for a more accurate understanding of the characteristics of elderly speech.

Acknowledgements

The authors wish to thank Mária Gósy and Zsófia Koren-Dienes for their help in preparing this paper.

References

de Andrade, C. R. F. & V. de Oliveira Martins. 2010. Speech fluency variation in elderly. *Pró-Fono Revista de Atualização Científica* 22(1): 13–18. https://doi.org/10.1590/S0104-56872010000100004

Beke, A., M. Gósy, V. Horváth, D. Gyarmathy & T. Neuberger. 2014. Disfluencies in Spontaneous Narratives and Conversations in Hungarian. In: . Fuchs, M. Grice, A. Hermes, L. Lancia & D. Mücke (eds.), *Proceedings of the 10th International Seminar on Speech Production (ISSP)*, 5–8 May 2014, Cologne, Germany, 29–32.

Boersma, P. & D. Weenink. 2008. Praat: Doing phonetics by computer (version 5.0.1). http://www.praat.org/ (accessed 28 October 2008).

Bóna, J. 2014. Temporal characteristics of speech: The effect of age and speech style. *Journal of the Acoustical Society of America* 136(2): EL116–EL121. https://doi.org/10.1121/1.4885482

Bóna, J. 2018. Non-verbal vocalizations in spontaneous speech: The effect of age. *Phonetician*, 115: 23–35.

Burke, D. M., D. G. MacKay, J. S. Worthley & E. Wade. 1991. On the tip of the tongue: What causes word finding failures in young and older adults. *Journal of Memory and Language* 30(5): 542–579. https://doi.org/10.1016/0749-596X(91)90026-G

Das, S., N. Gandhi, T. Naik & R. Shilkrot. 2019. Increase Apparent Public Speaking Fluency by Speech Augmentation. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12–17 May 2019, Brighton, UK, 6890–6894. https://doi.org/10.1109/ICASSP.2019.8682937

Duchin, S. W. & E. D. Mysak. 1987. Disfluency and rate characteristics of young adult, middle-aged, and older males. *Journal of Communication Disorders* 20(3): 245–257. https://doi.org/10.1016/0021-9924(87)90022-0

Fletcher, J. 2010. The prosody of speech: Timing and rhythm. In: W. J. Hardcastle, J. Laver & F. E. Gibbon (eds.), *The Handbook of Phonetic Sciences, Second Edition*, 521–602. Chichester, UK: Blackwell. https://doi.org/10.1002/9781444317251.ch15

Goberman, A. M., S. Hughes & T. Haydock. 2011. Acoustic characteristics of public speaking: Anxiety and practice effects. *Speech communication*, 53(6): 867–876. https://doi.org/10.1016/j.specom.2011.02.005

Gyarmathy, D. 2019. A néma szünetek és a hallható levegővétel viszonya a spontán beszédben [The relationship between silent pauses and audible breathing in spontaneous speech]. *Beszédkutatás* 27: 154–186.

Hnath-Chisolm, T., J. F. Willott & J. J. Lister. 2003. The aging auditory system: anatomic and physiologic changes and implications for rehabilitation. *International Journal of Audiology* 42(S2): 3–10. https://doi.org/10.3109/14992020309074637

Roberts, P. M., A. Meltzer & J. Wilding. 2009. Disfluencies in non-stuttering adults across sample lengths and topics. *Journal of communication disorders* 42(6): 414–427. https://doi.org/10.1016/j.jcomdis.2009.06.001

Searl, J. P., R. M. Gabel & S. J. Fulks. 2002. Speech disfluency in centenarians. *Journal of Communication Disorders* 35(5): 383–392. https://doi.org/10.1016/S0021-9924(02)00084-9

Torre, P. & J. A. Barlow. 2009. Age-related changes in acoustic characteristics of adult speech. *Journal of Communication Disorders* 42(5): 324–333. https://doi.org/10.1016/j.jcomdis.2009.03.001

Yairi, E. & N. F. Clifton. 1972. Disfluent speech behavior of preschool children, high school seniors, and geriatric persons. *Journal of Speech and Hearing Research* 15(4): 714–719. https://doi.org/10.1044/jshr.1504.714

Error type disfluencies in consecutively interpreted and spontaneous monolingual Hungarian speech

Maria Bakti
Department of Modern Languages and Cultures, University of Szeged, Szeged, Hungary

Abstract

Interpreting can be considered as a form of spontaneous speech, the key differences being that language change is involved in interpreting and the fact that speech production is influenced by several constraints during interpreting. Research has shown that the interpreting task influences the disfluency patterns of target language texts. The aim of this paper is to investigate how the frequency and distribution of error type disfluencies changes in the target language output of trainee interpreters as they progress in their training. Results indicate that there is no considerable change in the frequency and proportion of error type disfluencies in the target language texts recorded at the end of the second, third and fourth semesters of interpreter training. The proportion of error type disfluencies is higher in the consecutively interpreted texts than in the spontaneous monolingual speech of the students. This suggests that the complexity of the task, rather than progress in training, determines the disfluency pattern of consecutively interpreted target language texts.

Introduction

Speech production during interpreting has received considerable research attention to date. Simultaneous Interpreting (SI) can be considered as a variety of spontaneous speech (Goldman-Eisler, 1968); Kopczyński concurs with this view, highlighting that the target language (TL) text produced by an interpreter is "produced on the spot, on the basis of a previously unknown text" (Kopczyński, 1982: 257). This TL text is produced under time pressure and other cognitive constraints, and thus it can be expected to be prone to speech errors (Bakti, 2015: 368). Interpreters render a source language (SL) message in the TL, in other words they reproduce the ideas of the SL speaker instead of their own. In addition, they work based on an incomplete SL input.

Studies on fluency and pauses in interpreting have mostly focused on the simultaneous mode and include work on pauses (Tissi, 2000) and self-repairs (Petite, 2005). The analysis of error type disfluencies (ETDs) in simultaneously interpreted Hungarian TL texts showed that restarts, grammar errors and false words had the highest proportion in the TL output of interpreters, (Bakti, 2009) and the frequency of ETDs (ETDs/100 words) was between 2.8 and 6.2 in the case of trainee interpreters (Bakti, 2013).

However, the psycholinguistic aspects of Consecutive Interpreting (CI) have received limited research attention to date. CI comprises of two stages; the first is listening and note taking, during which active analysis takes place, followed by the production and note-reading stage. During this second stage, interpreters produce a target language text based on their memory and their notes, which can be either language dependent or may comprise of language-independent symbols, or both.

The basic tenets of Gile's Effort Models in interpretation are that interpretation requires mental energy that is available in limited supply, and that interpretation takes up almost all of this mental energy (Gile, 1995: 161). In CI, mental energy is used for the following Efforts during the listening and note-taking phase: listening and analysis, note-taking, short-term memory operations, and coordination. During the production phase, the following efforts require mental energy: remembering, note-reading, and production (Gile, 1995: 179). As competence develops during training, trainee interpreters learn how to best use their mental energy or attentional resources in order to produce high quality TL texts during CI. During this process of expert skill acquisition, some of the processes involved in CI become automatic (Albl-Mikasa, 2013).

Mead examined the control of pauses by trainee and professional interpreters in their A (L1) and B (L2) languages in CI (Mead, 2000, 2002) and found that the proportion of pauses was higher in the output of trainee interpreters when they were interpreting into their B language. In his second investigation Mead found that with the increase in interpreting experience, the proportion of hesitations related to grammatical and lexical problems decreased. Bakti and Bóna (2017) compared the disfluency patterns of spontaneous, semi-spontaneous, consecutively interpreted and sight translated texts. They have found that disfluencies are a function of the interpreting task; this finding is based on a cross-sectional study.

The aim of this paper is to investigate error-type disfluencies in consecutively interpreted Hungarian

https://doi.org/10.21862/diss-09-019-bakti

TL texts that were recorded in the framework of a longitudinal study with the aim to compare the proportion and frequency of the occurrence of ETDs in the TL texts recorded at different stages of a Master's Program in interpreting. In addition, figures are compared to the Hungarian monolingual spontaneous (i.e. not interpreted) speech of the participants.

This paper aims to answer the following research questions:

1. What is the proportion of ETDs in the consecutively interpreted (English to Hungarian) TL output of trainee interpreters recorded at different stages of their MA training?
2. How does the proportion of ETDs in interpreted texts differs from the proportion of ETDs in the spontaneous Hungarian monolingual speech of the students?
3. What are the most frequent disfluencies in the interpreted texts and how do they change as students progress in training?

Based on the literature I worked with the following hypotheses:

1. The proportion of ETDs will decrease in the interpreted texts as training progresses, because students will gain expertise and their TL speech production will become more fluent.
2. There will be more ETDs in the interpreted texts than in the spontaneous monolingual speech of the trainee interpreters, as speech production during CI is more complex than spontaneous monolingual speech production.
3. Even though students progress in training and gain expertise in interpreting, there will be no changes in the disfluency pattern of the TL consecutively interpreted texts, as disfluency patterns are task-specific.

Procedure

Five female and two male MA students of interpreting participated in the longitudinal study. Participation in the study was voluntary. Recordings were made of the students' consecutive interpreting, sight translation, spontaneous and semi-spontaneous speech production. The students' mean age was 23.3 years at the end of the second semester of their studies, when the first recordings were made. These were followed by recordings at the end of the 3rd and 4th semesters. The Hungarian monolingual spontaneous speech samples were recorded at the end of the 2nd semester. The A language or L1 of the students is Hungarian, their B language or active language is English (for 3 students) and Spanish (4 students). The C or passive language of the students was English (4 students), Italian (1 student), German (1), and French (1).

Even though English was B language for 3 students and C language for 4 students, their English background knowledge can be considered similar; students with English as their B language had been learning English for 14.3 years when the first recordings were made, and students with English as their C language had been learning English for 16 years at the time when the first recordings were made.

During the CI tasks at the end of the 2nd, 3rd and 4th semesters of their studies, students interpreted English source language texts of comparable length and lexical and syntactical complexity into Hungarian. No SL texts were re-used. All three SL texts were about similar topics: introducing an institution of higher education to prospective students. For further details of the SL texts, see Table 1.

Table 1. Source Language texts for the CI tasks.

	Semester 2	Semester 3	Semester 4
Topic	Williams College	Oxford University	Swansea University
Number of words SL	492	477	511
Average number of words TL texts	486.6	421.4	453.6
No. of sections	10	10	10

Some deviations from standard interpreting practice have to be noted. The recordings were made in a language laboratory without an audience present. These limitations undermine to some extent the ecological validity of the investigation.

The TL texts and the Hungarian monolingual spontaneous texts were transcribed and ETDs were identified in the texts, using the taxonomy of Gósy et al. (2009). The category of restarts was also added. Even though this category is a rather controversial one (Gyarmathy, 2015), as it can be seen both as an error or a disfluency rooted in uncertainty, it was included in the analysis to make it possible to compare the results with those of earlier investigations (Bakti, 2009). See Table 2 for the definitions and examples.

Results

First, the frequency of the occurrence of the ETDs was calculated. Results are shown in Table 3. The number of ETDs / 100 words of the target texts in consecutively interpreted texts show no considerable change in the course of the training. The frequency

Table 2. ETDs examined with definitions and examples.

ETD	Definition (Gósy et al., 2009; Gyarmathy, 2015)	Example from the Hungarian TL texts (student, semester)
fasle word	instead of the appropriate word, there is a different word in the surface structure	írták át vagy rajzolták át (WW4)
grammar error	a morphological or syntactical structure that is inconsistent with the norm	amely walesi Swanseaban található (E4)
blend	blend of two signs (word, phrase)	most pedig anélkül, hogy, a teljesség igénye nélkül (S4)
false start	articulation of a sound or string of speech sounds which do not constitute a word	lehetőségük van a diákoknak els_tanulni (L4)
TOT	knowledge of the morphology of the intended word, inhibition of the articulation of the phonetic structure	gazdasággal összefüggő, gazdasággal összefüggésben programokra (WW4)
ordering problems	perseveration, anticipation, metathesis	téz évben, tíz évben (E2)
slip	error in articulation, adding, replacing or deleting speech sounds	matekailag, matematikailag (E2)
restart	pronunciation of the activated and partly uttered word	lássunk néhány_néhányat (WW4)
multiple cause	errors that fit into more than one of the above categories	működik a az oktatás (M4) This could be seen both as a restart or a corrected grammar error.

Table 3. The frequency of ETDs in the output of trainee interpreters.

	2nd semester	3rd semester	4th semester	spontaneous speech
ETDs / 100 words	2.17	2.44	2.45	1.23
maximum	3.75 (Student L)	3.48 (Student E)	3.18 (Student Mi)	1.95 (Student Mi)
minimum	1.1 (Student S)	1.96 (Student Mi)	1.75 (Student E)	0.6 (Student WW)
standard deviation	0.879	0.548	0.515	0.507

of ETDs in the spontaneous monolingual speech of the students is lower than in the interpreted texts, illustrating that spontaneous speech production is less complex than the concurrent tasks of consecutive interpreting.

These averages however, hide considerable individual differences. At the end of the 2nd semester, Student L had 3.75 ETDs / 100 words, and the Student S had the lowest frequency of occurrence of ETDs: 1.1 ETDs / 100 words. At the end of the third semester, Student E had 3.48 ETDs / 100 words, and Student Mi 1.96. At the end of the 4th semester Student Mi had 3.18 ETDs / 100 words, and Student E 1.75. In spontaneous speech, the lowest figure was 0.6 ETDs / 100 words, and the highest 1.95 ETDs / 100 words.

Second, the proportion of ETDs was calculated.

The results are shown in Figure 1. In the 2nd and 4th semester CI tasks, the proportion of grammar errors was the highest. In the 2nd semester, this was followed by the categories of false word and restarts. In the recordings made after the 3rd semester, the proportion of multiple cause ETDs was highest, followed by grammar errors, restarts ranked third, followed by false words. In the recordings made after the 4th semester, the category of false words followed grammar errors, and restarts ranked third. In the spontaneous monolingual speech production of the students, grammar errors and restarts had the same proportion, followed by false starts.

Summary

In summary it can be stated that the examined ETDs were more frequent in the consecutively interpreted Hungarian texts than in the spontaneous monolingual Hungarian speech of the trainees. This supports the finding of Bakti and Bóna (2017) in that the complexity of the CI task is mirrored in the disfluencies; the more complex the task, the more frequent the disfluencies are.

The results show that the average frequency of ETDs does not change considerably as students progress in their training, however, considerable individual differences exist. This finding does not confirm the first hypothesis.

The second hypothesis, that there would be more ETDs in the interpreted than in the spontaneous monolingual speech of the students, was confirmed by the data.

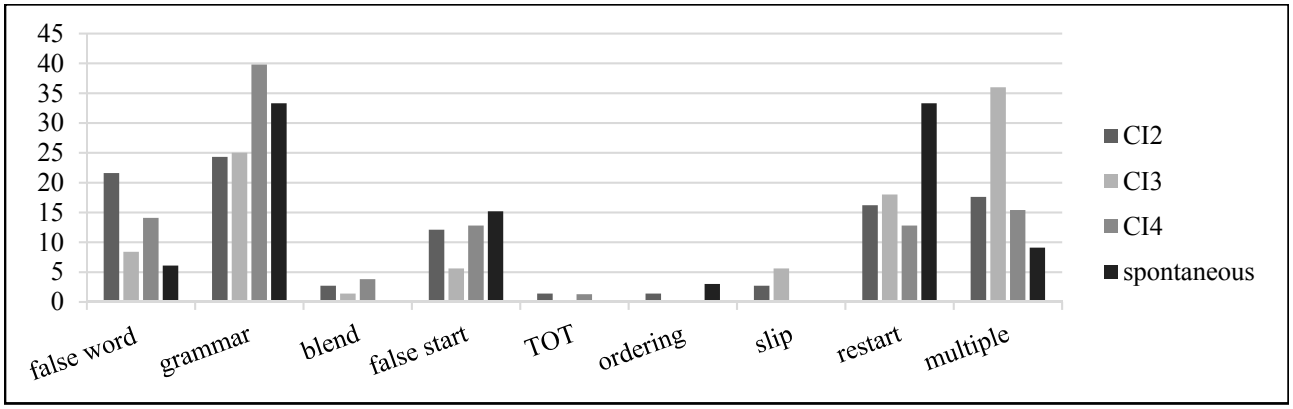


Figure 1. The proportion of ETDS in the output of trainee interpreters.

The distribution of ETDs was similar in the consecutively interpreted TL texts, which supports the third hypothesis.

Limitations of this investigation should also be noted, namely the small number of participants. The results should be tested against larger corpora.

References

Albl-Mikasa, M. 2013. Developing and cultivating expert interpreter competence. *The Interpreters' Newsletter* 18: 17–34.

Bakti, M. 2009. Speech Disfluencies in Simultaneous Interpreting. In: D. De Crom (ed.) *Selected Papers of the CETRA Research Seminar in Translation Studies* 2008, https://www.arts.kuleuven.be/cetra/papers/files/bakti.pdf. Published online by CETRA, KU Leuven, 2009.

Bakti, M. 2013. Théorie du sens revisited. CLI in the target language output of simultaneous interpreters. In: B. Lewandowska-Tomaszczyk & M. Thelen (eds.) *Translation and Meaning Part 10*, 363–370. Maastricht: Zuyd University of Applied Sciences.

Bakti, M. 2015. Slips. In: F. Pöchhacker (ed.) *The Routledge Encyclopedia of Interpreting Studies*, 386–387. New York, London: Routledge.

Bakti, M. & J. Bóna. 2017. A contrastive analysis of disfluency markers in four different settings. Conference poster: Fluency and Disfluency Across Languages and Language Varieties Conference, Université Catholique de Louvain, Louvain-le-Neuve, Belgium, February 2017.

Gyarmathy, D. 2015. Diszharmóniás jelenségek, megakadások a beszédben [Disharmonies and speech disfluencies]. In: M. Gósy (ed.) *Diszharmóniás jelenségek a beszédben* [Disharmonic phenomena in speech], 9–47. Budapest: MTA Nyelvtudományi Intézet.

Gile, D. 1995. *Basic Concepts and Models for Interpreter and Translator Training*. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.8(1st)

Goldman-Eisler, F. 1968. *Psycholinguistics; experiments in spontaneous speech*. London, UK: Academic Press.

Gósy M., J. Bóna, T. Gráczy, V. Horváth, A. Imre, & T. Neuberger. 2009. Nyelvbotlás korpusz, 6. Rész [Speech Error Corpus, Part 6]. *Beszédkutatás* 2009: 257–267.

Kopczyński, A. 1982. Effects of some characteristics of impromptu speech on conference interpreting. In: N. Enkvist (ed.) *Impromptu Speech: A Symposium*., 255–266. Abo: Abo Akademie.

Mead, P. 2000. Control of pauses by trainee interpreters in their A and B languages. *The Interpreters' Newsletter* 10, 89–102.

Mead, P. 2002. Exploring hesitation in consecutive interpreting: An empirical study. In: G. Garzone & M. Viezzi (eds.), *Interpreting in the 21st Century. Challenges and Opportunities*, 73–82 Amsterdam: John Benjamins. https://doi.org/10.1075/btl.43.08mea

Petite, C. 2005. Evidence of repair mechanisms in simultaneous interpreting: A corpus-based analysis. *Interpreting* 7(1): 27–49. https://doi.org/10.1075/intp.7.1.03pet

Tissi, B. 2000. Silent pauses and disfluencies in simultaneous interpretation: A descriptive analysis. *The Interpreters' Newsletter* 10, 103–127.

Effects of speech rate changes on pausing and disfluencies in cluttering

Johanna Pap
Department of Applied Linguistics and Phonetics, ELTE Eötvös Loránd University, Budapest, Hungary

Abstract

People with cluttering (PWC) often receive feedback, such as "Slow down!", even so, this fluency disorder cannot be cured by only slowing down the speakers' speech rate. When PWC accelerate their speech rate, language planning difficulties and word structure errors might occur, which might result in breakdowns in fluency and/or intelligibility. In the present paper characteristics of the frequency of disfluencies were examined in four different speech tasks from deliberately slow to maximum speech rate, whether speech rate changes have effects on cluttered speech. Twenty participants of this investigation were individuals suspected of cluttering with ages between 20 and 50 years of both genders. The results show that PWC are able to change, not only their speech rate but articulatory rate as well. Moreover, disfluencies were produced the most frequently in the speech task of maximum speech rate, where PWC do not have enough time for speech planning. The research provides empirical, measured data for a better insight into the nature of cluttering. Understanding the correlation between speech rate and disfluencies in cluttered speech is fundamental to improve the diagnosis of cluttering.

Introduction

Due to its heterogeneous nature, cluttering has been defined in many ways over the years. There is no agreement on the concepts which dimensions are affected if it is speech, language, communication-based disorder or syndrome (Ward, 2006, St Louis et al., 2007, 2011, Van Zaalen-Op't Hof, Wijnen & De Jonckere, 2009, Van Zaalen & Reichel, 2015). However, all definitions of cluttering available in the literature consider abnormal fluency as an obligatory symptom of the disorder. According to the most widely accepted working definition, cluttering is a fluency disorder characterized by too fast and/or irregular speech rate perceived by listeners. To diagnose cluttering "rapid and/or irregular speech rate must further be accompanied by one or more of the following symptoms: a) excessive 'normal' disfluencies; b) excessive collapsing or deletion of syllables; and/or c) abnormal pauses, syllable stress, or speech rhythm." (St. Louis & Schulte, 2011: 241–242).

Disfluencies or word structure errors might occur when people with cluttering (PWC) execute speech at a fast or irregular rate (Daly, 1992; Van Zaalen, 2009; Bóna, 2010). Van Zaalen (2009) emphasizes that the frequency of symptoms will be increased when PWC do not have enough time for linguistic planning.

As excessive disfluencies might be one of the key symptoms in cluttering, recently, the number of studies investigating the nature of disfluencies in cluttered speech has been increasing. The findings though show contradictory results. De Oliveira et al., (2013) found more disfluencies produced in cluttered speech than in typical speech, in addition, a different fluency profile was examined in relation to normal and stuttering-like disfluencies (De Oliveira et al., 2010). Further, Garnett and St. Louis (2010) also reported twice as many disfluencies in PWC's speech than in the speech of typical speakers. Several studies, however, found no significant differences between PWC and typical speakers in the frequency of all disfluencies (Bakker et al., 2011; Myers et al., 2012; Bóna, 2016, 2018). Contradictory results might emerge due to the small number of participants and large individual characteristics.

To the author's knowledge, speaking rate changes in cluttered speech have been scarcely investigated from the point of view of disfluencies. Although some studies showed possible relationships between speech rate and disfluencies in spontaneous speech. Bóna (2012) studied the speech of a clutterer in four different speaking styles and the results confirmed earlier findings, that more pauses but fewer other disfluencies and phonological inaccuracies occurred when the clutterer slowed down his speech. As reported by De Oliveira et al. (2013), a greater frequency of typical disfluencies was produced as speech rate increased.

Nevertheless, there is still a relevant question to be addressed: what happens when PWC change their speech pace on purpose? Therefore the aim of this study was to examine if the characteristics of the frequency of disfluencies depend on speaking rate changes in cluttered speech, moreover, explore the strategies of speech rate changes. It was hypothesized that 1) Most PWC are not able to change their speech rate. 2) Higher frequency of disfluencies will occur at maximum and even faster rate than at a self-determined normal and deliberately slow rate.

https://doi.org/10.21862/diss-09-020-pap

Methods

Subjects

Participants of this investigation were individuals suspected of cluttering with ages between 18 and 50 years of both genders (16 females and 4 males). The mean age for PWC was 28.95 years (*s.d.* = 7.15 years). PWC were recruited by two speech therapists after listening two times the recordings. The diagnostic decisions were based on their subjective impressions, the speech-language pathology anamnesis and the working definition of cluttering (St Louis & Schulte, 2011). All participants were native Hungarian speakers with normal hearing, coming from a similar social and cultural background. They did not have any co-occurring speech and language disorders, did not have a history of stuttering and never received speech therapy. All of them considered themselves to be clutterers and volunteered for the tasks.

Material

Participants were asked to describe the steps of making scrambled eggs in four different speech rate: a) self-determined comfortable (normal) speech rate, where the only instruction was to talk freely about the scrambled eggs-story; b) deliberately slow speech rate (slow); c) self-determined maximum (fast) speech rate, where PWC were instructed to speak as fast as they can; d) even faster speech rate, which refers to maximum values throughout this paper. Participants were only told to talk about the scrambled egg-story, they were not aware of the specific instructions followed by. The method of the four rate conditions based on a previous study, in which PWC were asked to produce diadochokinetic syllable trains and related real words (Bakker et al., 2011).

Procedures

Speech samples were recorded in a soundproof chamber and a quiet environment. Speech samples were transcribed using Praat (Boersma & Weenink 2008). Disfluencies (filled pauses, whole-word repetitions, part-word repetitions, phrase-repetitions, incomplete phrases, broken words, and revisions) were annotated and analyzed. Filler words and prolongations were not counted in this analysis as they based on subjective decisions, and the rate of agreement of the speech therapists was not 100% in all cases.

The following measurements were obtained: the ratio of pauses, frequency of occurrence of pauses, duration of pauses, speech rate, articulatory rate and frequency of occurrence of disfluencies. Frequency values show how many disfluencies and pauses occurred in 100 syllables.

All calculations and ratings were carried out twice by the author, two weeks apart. The results of the two analyses were similar in 100% of the cases. Statistical analyses (one-way ANOVA and repeated measures ANOVA, Pearson's correlation) were carried out by SPSS 21.0 software in a 95% confidence level.

Results

Speech rate and articulatory rate

Figure 1. shows the speech rate and articulatory rate in all speech tasks. The lowest values of speech pace were found in the task of deliberately slow speech rate. The average speech rate was 3.65 syllables/s (*s.d.* = 0.71), the average articulatory rate was 5.07 syllables/s (*s.d.* = 0.84). On the other hand, the highest values of speech pace were found in the task of maximum speech rate, where PWC produced two more syllables/s. The average speech rate was 5.71 (*s.d.* = 1.09), the average articulatory rate was 7.16 (*s.d.* = 1.10). Significant differences were found between all tasks except self-determined normal and fast speech rates. The results of the statistical analysis are shown in Table 1.

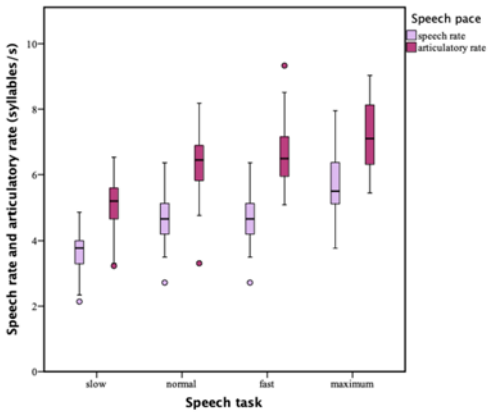


Figure 1. Values of speech rate and articulatory rate in all speech tasks

Characteristics of speech pauses

Examining the ratio of speech pauses, significant differences were found between normal and fast ($Z = 4.810$, $p = 0.043$), slow and fast ($Z = 7.578$, $p = 0.011$), slow and maximum speech rate ($Z = 7.613$, $p = 0.005$). Considering all participants, the average ratio of pauses did not show a large variation between speech tasks. Although the ratio of pauses increased as PWC slowed down their speech rate, a high standard deviation (8.09) at maximum speech pace indicates large individual characteristics.

During PWC were accelerating their speech rate, the duration of pauses started to decrease, except in the task of maximum speech rate, where a large standard deviation (0.24 s) was found. In the maximum speech rate the average duration of pauses (0.43 s) was similar to the values in normal speech rate (0.44 s). However, smaller standard deviation was

found in normal speech rate (0.08 s) than any other speech tasks. There were significant differences only between slow and fast ($Z = 0.137$, $p = 0.001$), normal and fast speech rate ($Z = 0.083$, $p = 0.012$).

Table 1. Results of the statistical analysis of values of speech pace

		slow		normal		fast		maximum	
		Z	*p*	*Z*	*p*	*Z*	*p*	*Z*	*p*
Speech rate	slow	-	-	0.996	0.002	3.920	0.000	−2.058	0.000
	normal	0.996	0.002	-	-	−2.856	0.004	−1.062	0.001
	fast	−3.920	0.000	−2.856	0.004	-	-	−1.923	0.055
	maximum	−2.058	0.000	−1.062	0.001	−1.923	0.055	-	-
Articulatory rate	slow	-	-	1.212	0.004	−1.549	0.000	−2.094	0.000
	normal	1.212	0.004	-	-	−0.337	1.000	−0.883	0.022
	fast	−1.549	0.000	−0.337	1.000	-	-	−0.546	0.005
	maximum	−2.094	0.000	−0.883	0.022	−0.546	0.005	-	-

The frequency of occurrence of pauses was analyzed too, values were calculated per 100 syllables. The results, as shown in Figure 3., indicate that the frequency of speech pauses was influenced by speech rate changes as the frequency decreased due to slower speech rates. Large individual characteristics were found in all speech tasks, especially in the slow (4.31) and maximum speech rate (4.41). Further, there were significant differences between all speech tasks except between normal and fast speech rate (Table 2.).

Frequency of occurrence of disfluencies

The frequency of disfluencies are highlighted in Figure 4. The results show that disfluencies were produced the most frequently in the maximum speech rate, where the mean frequency of disfluencies was 5.19 (*s.d.* = 3.78). More than half of the PWC produced fewer disfluencies than the mean frequency in the data. However, two PWC, who also used interjections and commented on their speech, produced more than 10 disfluencies in 100 syllables.

Table 2. Results of the statistical analysis of frequency of pauses.

	slow		normal		fast		maximum	
	Z	*p*	*Z*	*p*	*Z*	*p*	*Z*	*p*
slow	-	-	−2.688	0.007	5.083	0.001	−7.123	0.000
normal	−2.688	0.007	-	-	−1.456	0.145	−2.725	0.006
fast	5.083	0.001	−1.456	0.145	-	-	2.040	0.006
maximum	−7.123	0.000	−2.725	0.006	2.040	0.050	-	-

Comparing all speech tasks significant differences were found between the maximum and all other types of speech rate changing tasks in the frequency of occurrence of disfluencies ($F = 4{,}797$, $p = 0.021$). Stronger significant difference was found between maximum and slow, maximum and fast speech rate ($Z = 1.654$, $p = 0.007$), than the significant difference between maximum and normal speech rate ($Z = 1.216$, $p = 0.39$).

Correlation between the frequency of occurrence of disfluencies and speech rate changes were analyzed, as well. There was a negative correlation in the task of normal ($r = -0.633$, $n = 20$, $p = 0.003$) and fast speech rate ($r = -0.618$, $n = 20$, $p = 0.004$), which indicates that PWC produced less disfluencies when the speech rate was accelerated.

Discussion

This paper analyzed the frequency of disfluencies and pauses in cluttered speech in four different speech rates. It was hypothesized that PWC are not able to change their speech rate (1st hypothesis). However, results did not confirm the first hypothesis, as PWC were able to change not only their speech rate but their articulatory rate in all tasks as well. In conclusion, PWC are capable of changing the duration of speech sounds.

As the speech rate increased the ratio and mean duration of pauses was decreased except in the maximum rate, where a high standard deviation was found. It means that the task of maximum rate change has different effects on PWC.

It was assumed that a higher frequency of disfluencies will occur as PWC increase their speech rate (2nd

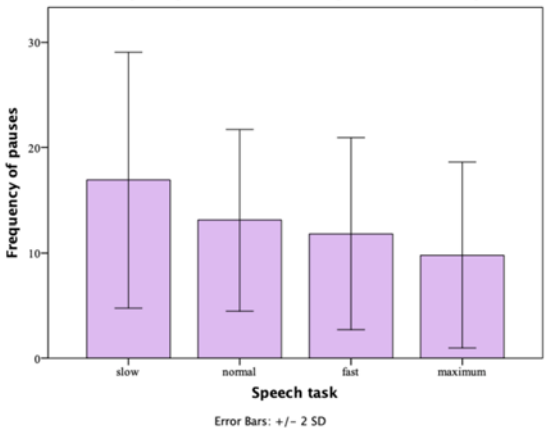


Figure 2. Mean frequency of occurrence of pauses in 100 syllables

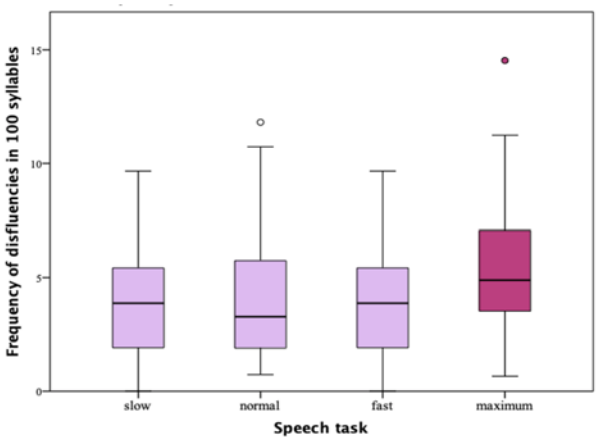


Figure 3. Values of speech rate and articulatory rate in all speech tasks

hypothesis). Disfluencies were produced the most frequently in the task of maximum speech rate which indicates that lack of time influences language planning difficulties in cluttered speech. Similarly to these results, De Oliveira et al. (2013) found higher frequency of disfluencies as speech rate increased. The results also confirm the theory of Van Zaalen (2009) that PWC are not capable of adequately adjust their speech rate to the demands of the speaking situation.

The second hypothesis was partially correct, as the frequency of disfluencies produced in fast speech rate were similar to the frequency of disfluencies in normal speech rate. We must not disregard that the present study had some limitations. One of them can be found in the methods since the task of the slow speech rate was followed by the task of the fast speech rate. Comparing slow and fast speech rate PWC were able to change their speech rate, but this rate change reached only the values of the normal speech rate.

Although several earlier studies reported that disfluencies occur more in high articulatory rate, the present paper provide useful additional evidence for future studies. According to the results the task of the maximum speech rate affects more the frequency of disfluencies than the actual articulatory rate. PWC produced more disfluencies when they did not have enough time for language planning at maximum rate even though they have already described the scrambled eggs-story three times before. The analysis leads to the conclusion that psychological stress might have a strong influence on speech efficiency in cluttered speech.

Nonetheless, a relatively small number of PWC participated, the results might provide useful information to the diagnosis of cluttering. Additional future research directions include comparing the efficiency and characteristics of strategies of speech rate changing tasks of PWC and typical speakers, in addition the affects of psychological stress on cluttered speech.

Acknowledgements

This research was supported by the ÚNKP-18-3 New National Excellence Program of the Ministry of Human Capacities.

References

Bakker, K., F. L. Myers, L. J. Raphael & K. O. St. Louis. 2011. A preliminary comparison of speech rate, self-evaluation, and disfluency of people who speak exceptionally fast, clutter, or speak normally. In: D. Ward & K. S. Scott (eds.): *Cluttering: A Handbook of Research, intervention and education*, 45–65. East Sussex, UK: Psychology Press.

Boersma, P. & D. Weenink. 2008. Praat: Doing phonetics by computer (version 6.0.46). http://www.praat.org/ (accessed 8 January 2019).

Bóna, J. 2018. Clustering of disfluencies in typical, fast and cluttered speech. *Clinical Linguistics & Phonetics* 33(5): 393–405. https://doi.org/10.1080/02699206.2018.1513075

Bóna, J. 2016. Characteristics of pausing in normal, fast and cluttered speech. *Clinical Linguistics & Phonetics* 30(11): 888–898. https://doi.org/10.1080/02699206.2016.1188421

Bóna, J. 2012. Linguistic-phonetic characteristics of cluttering across different speaking styles: A pilot study from Hungarian. *Poznań Studies in Contemporary Linguistics* 48: 203–222. https://doi.org/10.1515/psicl-2012-0010

Bóna, J. 2010. Mindig hadar-e a hadaró? – Akusztikai-fonetikai vizsgálatok tanulságai. Gyógypedagógiai szemle [Do people with cluttering always clutter? – Lessons from acoustic-phonetic studies]. *A Magyar Gyógypedagógusok Egyesületének Folyóirata* 38(1): 24–31.

Daly, D. 1992. Helping the clutterer: Therapy consideration. In: F. Myers F. & K. St. Loui (eds.): *Cluttering: A Clinical Perspective*. Leicester, UK: Far Communications. (Reissued in 1996 by San Diego, CA: Singular), 107–124.

De Oliveira, C. M. C., A. P. L. Bernardes, G. A. F. Broglio & S. A. Capellini. 2010. Perfil da fluência de indivíduos com taquifemia [Speech fluency profile in cluttering individuals]. *Pró-Fono Revista de Atualização Científica* 22(4): 445–50. https://doi.org/10.1590/S0104-56872010000400014

De Oliveira, C. M. C., G. A. F. Broglio, A. P. L. Bernardes, & S. A. Capellini. 2013. Relationship between speech rate and speech disruption in cluttering. *CoDAS: Publication of the Brazilian Society of Speech-Language Pathology and Audiology* 25(1): 59–63. https://doi.org/10.1590/S2317-17822013000100011

Garnett, E. O. & St. Louis, K. O. 2010. Hesitations in cluttered speech. Paper presented at the *1st Online Conference on Cluttering*. Retrieved from: http://www.mnsu.edu/comdis/ica1/papers/garnett2c.html

Myers, F. L., K. Bakker., K. O. St Louis, & L. J. Raphael. 2012. Disfluencies in cluttered speech. *Journal of Fluency Disorders* 37(1): 9–19. https://doi.org/10.1016/j.jfludis.2011.10.001

St. Louis, K. O., Myers, F. L., Bakker, K. & Raphael., L. J. 2007. Understanding and treating cluttering. In: E. G. Conture & R. F. Curlee (eds.): *Stuttering and related disorders of fluency, 3rd Edition*, 297–325. New York: Thieme.

St. Louis, K. O. & K. Schulte. 2011. Defining cluttering: The lowest common denominator. In: D. Ward & K. S. Scott (eds.): *Cluttering: A Handbook of Research, intervention, education*. East Sussex, UK: Psychology Press, 233–253.

Van Zaalen, Y. & I. K. Reichel. 2015. *Cluttering: Current views on its nature, diagnosis, and treatment*. Bloomington, US: iUniverse.

Van Zaalen-Op't Hof, Y., F. Wijnen & P. H. De Jonckere. 2009. Differential diagnostic characteristics between cluttering and stuttering – Part one. *Journal of Fluency Disorders* 34(3): 137–154. https://doi.org/10.1016/j.jfludis.2009.07.001

Van Zaalen, Y. 2009. *Cluttering identified. Differential diagnostics between cluttering, stuttering and speech impairment related to learning disability*. Ph.D. dissertation, University of Utrecht, The Netherlands.

Ward, D. 2006. *Stuttering and cluttering. Frameworks for understanding and treatment*. East Sussex: Psychology Press.

Disfluencies in mildly intellectually disabled young adults' spontaneous speech

Julianna Jankovics [1] *and Luca Garai* [2]
[1]*Doctoral School of Linguistics, Applied Linguistics, Eötvös Loránd University, Budapest, Hungary*
[2]*Department of English Linguistics, Eötvös Loránd University, Budapest, Hungary*

Abstract

The study analyzes various hesitations and repairs in the spontaneous speech of mildly intellectually disabled women. The main research questions of the study focus on the similarities and differences in the frequency of disfluencies and the duration of pauses between the spontaneous speech of mildly intellectually disabled and mentally healthy young adults. Our results show that hesitation phenomena were more frequent among intellectually disabled subjects in spontaneous speech, while repairs occurred more frequently among control subjects in guided spontaneous speech.

Introduction

The definition of disfluency is not consistent (Lickley, 2015), as the phenomena occurring in spontaneous speech are investigated by many fields of science (Roberts, Meltzer & Wilding, 2009). According to Lickley (2015: 451–452), who also notes that there is no universal definition for the term, disfluency occurs when a speaker fails to convey their message without any error, correction, or struggle to express themselves, but he also states that perfect fluency is almost impossible for most speakers and that "everyone is disfluent some of the time."

The classification of disfluency types is also not consistent in the literature. Those analyzing typical speech usually divide these phenomena in two groups, hesitations and repair-type disfluencies. While the former occur when the speaker is unsure of what they want to say yet and uses these disfluencies as devices to stall for time, the latter are present when the speaker's message is not realized correctly or when the rules of the language are breached (cf. Lickley, 2015). Clinical literature differentiates between typical speech and fluency disorders based on the frequency and types of disfluencies. The latter classification mainly encompasses the analysis of cluttering and stuttering (cf. Bóna, 2018). These works make a distinction between typical and stuttering-like disfluencies.

Numerous studies have focused on these phenomena occurring in the speech of people with mental disabilities. Rossi et al. (2011) investigated the speech of subjects with Williams Syndrome and found that disfluencies, particularly hesitations, repetitions, and pauses, were significantly more frequent in their speech than in that of their control subjects. Coppens-Hofman et al. (2013) studied the spontaneous speech of adults with mild or moderate intellectual disabilities and their results found typical dysfluency patterns which feature cluttering much more prominently when compared to stuttering.

The current research focuses on disfluencies in the speech of mildly intellectually disabled subjects (with an IQ score between 50 and 69 points, ICD-11, 2018) and compares them to those in the speech of mentally healthy subjects. The study follows the classification outlined by Lickley (2015), which differentiates between two disfluency types, hesitations (including prolongation, silent and filled pauses, and repetition) and repairs (including substitution, insertion, and deletion). All phenomena are included in the current analysis, except for prolongation.

The study aims to discover how frequently the various disfluency types occur in two kinds of spontaneous speech and what differences there are between the disfluencies of female subjects with and without mental disabilities. Based on the literature, the following hypotheses were outlined:

(1) mildly intellectually disabled subjects exhibit more hesitations in both speech types when compared to control subjects;
(2) the spontaneous speech of intellectually disabled people contains more disfluencies in total;
(3) filled and silent pauses in the speech of intellectually disabled subjects are longer in duration.

Subjects, material, and method

The subjects of the current research were 10 mildly intellectually disabled women between the ages of 17.8 and 21.7 years (ID1–ID10) and 10 mentally healthy women (control subjects) of similar ages, between 18.1 and 29.1 years (C1–C10). All participants were native speakers of Hungarian, spoke the standard vernacular, and were non-smokers. The speech samples were all recorded in a quiet environment. In total, four types of speech were recorded, of which two are subject to analysis in the current paper. In the Interview, the subjects were asked about school or work, free time activities, and

https://doi.org/10.21862/diss-09-021-jank-gara

family by the recording operator. During the Picture Story, subjects had to tell a story using a coherent set of eight black-and-white images. In both exercises, subjects could speak freely, with no time limit. These exercises were partially identical to those found in GABI Child Language and Speech Database (Bóna et al., 2014) and Andrea Tóth's (2017) research.

Recordings were annotated on the segmental level using *Praat 6.0.43* (Boersma & Weenink, 2018). Disfluencies were marked by hand in all speech samples. The disfluencies analyzed in the current paper are the following: silent pauses, filled pauses, repetitions, as well as substitutions, insertions, and deletions. These disfluency types underwent frequency analysis, and in the case of silent and filled pauses, a duration analysis was conducted using a script. In Interviews, the first 300 syllables were taken into account, while the first 100 syllables of each Picture Story were analyzed. For the purposes of comparability, in cases where the syllable count in the speech of mildly intellectually disabled subjects did not reach the desired limit of 300 and 100 syllables but was at least 70% of these, the results were proportionally scaled to match other samples. Table 1 shows the syllable count produced in the two speech types by ID subjects. Statistical analyses were not conducted due to the relatively small number of subjects.

Table 1. Number of syllables in the full samples in the mildly intellectually disabled group

Subject	Interview	Picture Story
ID1	106	60
ID2	> 300	> 100
ID3	> 300	> 100
ID4	> 300	> 100
ID5	> 300	74
ID6	233	98
ID7	> 300	56
ID8	130	> 100
ID9	> 300	> 100
ID10	> 300	> 100

Results

When analyzing the samples, the following subjects and recordings were not included in the calculation of results. Subject ID1 was excluded from the analysis due to the fact that her syllable count did not reach the predetermined limit in either the Interview or the Picture Story. Disfluencies in the Interview sample of subject ID8 and in the Picture Story sample of subject ID7 were also disregarded during the evaluation for similar reasons.

Frequency analysis of disfluencies

On average, ID subjects produced 9.3 filled pauses, 32.6 silent pauses, and 2.5 repetitions during the first 300 syllables of the Interview samples. Subjects ID6 and ID7 did not produce any repetitions. Subject ID4 had the fewest filled pauses (3), while subject ID9 had the most (21). There were even more substantial differences between individual subjects when it came to silent pauses. Only 13 instances were detected in the first 300 syllables of subject ID7, whereas ID4 had 65 silent pauses in total. When it came to repairs, there was an average of 0.25 substitutions, 0.5 insertions, and 0.13 deletions among ID subjects. There were only two substitutions in total, both from subject ID9. Insertions were found in the speech of subjects ID3, ID4, and ID9. Only one deletion was detected, which belonged to subject ID3.

Among mentally healthy subjects, these numbers were the following in the Interview samples. On average, filled pauses occurred 8.6 times, silent pauses 20.8 times, and repetitions 3.4 times. In the first 300 syllables, there was an average of 1.3 substitutions, 0.3 insertions, and 0.5 deletions. This means that there were more silent pauses and insertions in the mildly intellectually disabled group, while the other disfluencies were more prominent among control subjects. This can be explained by the fact that the speech production (and perception) of intellectually disabled people is influenced by the weakness of stimulus processing functions (cf. Lukács & Kas, 2014), and these people need more time to form their thoughts into words. Both silent and filled pauses, as well as insertion, are suitable for gaining some of the extra time needed.

In Picture Story samples, there was an average of 3.1 filled pauses, 18.0 silent pauses, and only 0.5 repetitions among ID subjects. Both substitutions and insertions occurred 0.25 times on average (with only two subjects producing them each), while the occurrence of deletions was 0.6 on average. It is important to note that while the frequency of occurring silent pauses was nearly identical in both sample types (close to 3.1 per 100 syllables), the frequency of filled pauses was much higher in Picture Story samples (18 per 100 syllables) than in Interview samples (10.9 per 100 syllables).

Picture Story samples of control subjects yielded the following results: in the 100 syllables analyzed, there was an average of 3.3 filled pauses, 10.4 silent pauses, and 0.1 repetitions and substitutions alike (one occurrence each, from C4 and C10). In this speech sample type, neither insertions nor deletions were found.

If we compare the frequency of hesitations and repairs in total, it can be said that the former phenomena occurred more frequently in both speech sample types. Hesitations were more frequent among ID subjects, whereas more instances of repairs were collected from control subjects in the Interview (Figure 1).

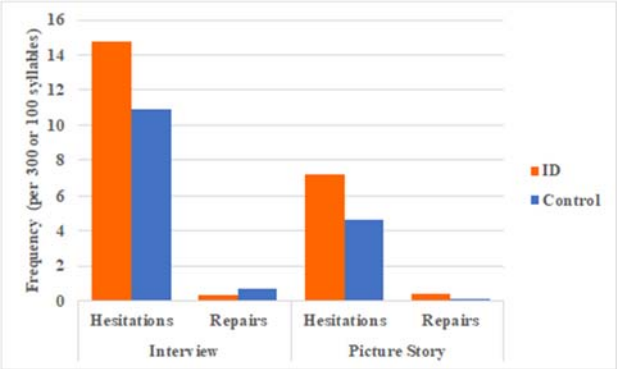


Figure 1. Average frequency of hesitations and repairs in the two speech types.

Hesitation and repair occurrences were analyzed collectively in both sample types. For the purposes of comparability, the total sum of disfluencies occurring in the speech of ID subjects was scaled up to make up for the two excluded samples in each speech type and match the sample size of control subjects. Table 2 shows the number of disfluencies detected in each complete set of samples after the scaling process.

Table 2. Total number of disfluencies produced by each group in both speech sample types.

	Interview	
Subjects	No. of syllables total	No. of disfluencies
ID	2400 (scaled to 3000)	452.4
Control	3000	343

	Picture Story	
Subjects	No. of syllables total	No. of disfluencies
ID	800 (scaled to 1000)	226.3
Control	1000	139

As it can be seen, ID subjects produced a higher number of disfluencies in both speech sample types, therefore their speech can be described as less fluent. However, the same calculations were done without the inclusion of silent pauses, in which case, the results are quite different between the two sample types. Figure 2 shows that the rate of disfluencies is reversed, as overall, ID subjects produced 126.5 disfluencies (excluding silent pauses) and control subjects produced 141. At the same time, 46.8 disfluencies were detected in all Picture Story ID samples, while only 35 disfluencies occurred in samples from control subjects. This corroborates that (at least among the subjects of the current research) silent pauses have a significant role in spontaneous speech. The high frequency of silent pauses among ID subjects might be explained by the fact that there is no need to pay attention to articulatory expressions during this type of hesitation, therefore it can be used to gain more time for speech planning.

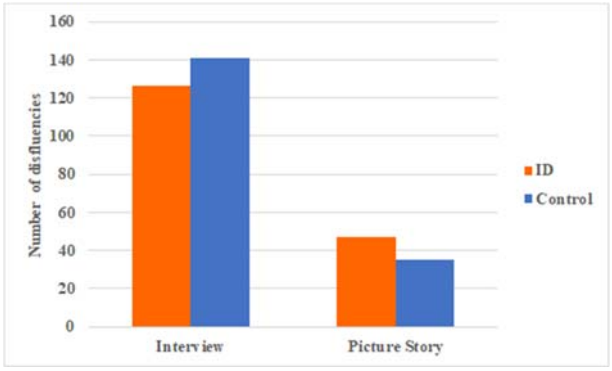


Figure 2. Number of disfluencies (without silent pauses).

Duration of filled and silent pauses

In the Interview, the filled pauses of ID subjects were 0.44 seconds long on average, while silent pauses lasted for 0.77 seconds. There was a large difference in the lowest and highest values between the two pause types. The shortest filled pause was 0.04 seconds long and the longest was 1.3 seconds, whereas the shortest silent pause was 0.06 seconds and the longest 5.8 seconds. As for control subjects, the average of filled pause lengths was 0.36 seconds, and silent pauses were 0.48 seconds long. These mean values, therefore, were higher among ID subjects.

There were no filled pauses in the Picture Story samples of subjects ID3, ID5, and ID10. Other ID subjects' filled pauses were 0.44 seconds long on average. The mean length of silent pauses was 0.87 seconds. Like in the Interview, the durations in the control subject group were shorter in Picture Story samples as well. Filled pauses were 0.38 seconds long, while silent pauses spanned an average 0.51 seconds.

If the pauses occurring in both speech types are taken into account, the average length of filled pauses from ID subjects was 0.44 seconds, while that of silent pauses was 0.80 seconds. The mean values for control subjects were 0.37 seconds (filled pauses) and 0.49 seconds (silent pauses). Figure 3 shows the length distribution of all recorded filled and silent pauses. It can be seen that silent pauses were lengthier in both subject groups, but as established before, the difference between filled and silent pauses was more prominent among ID subjects.

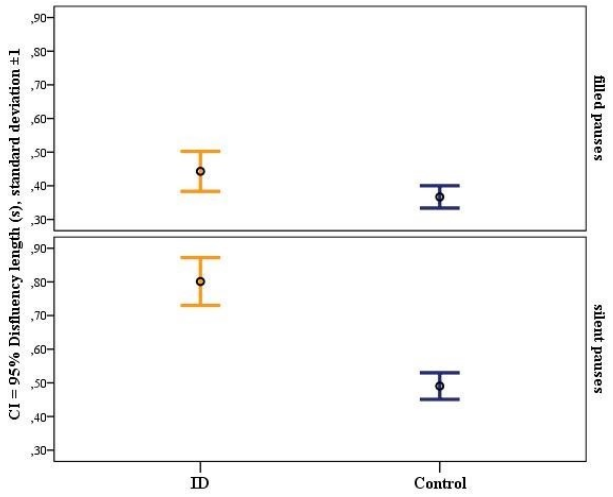


Figure 3. Length of filled and silent pauses in the two subject groups.

Discussion and conclusion

The first hypothesis was corroborated by our results, which can be traced back to numerous reasons. Hesitations include silent and filled pauses, which appeared more prominently among ID subjects. Hesitations allow the speaker to rethink their message and continue expressing their thoughts without inserting any new verbal elements.

Our findings also support the second hypothesis, as there were more disfluencies detected overall in both sample types of ID subjects when all six disfluency types were included, although there were meaningful individual differences. It is also important to note that when silent pauses were excluded from the analysis, it was shown that control subjects exhibited more disfluencies in the Interview. This discovery sheds light on the role silent pauses play in the spontaneous speech of intellectually disabled people. Repairs occurred sparingly in both sample types of ID subjects, which can be explained by the complexity of having to continue with the articulatory planning while simultaneously introducing new linguistic elements into the speech.

The third hypothesis, which presumed that the average duration of silent and filled pauses would be longer in the speech of ID subjects, was also corroborated. The reason behind this finding might be these subjects' difficulty in linguistic and speech performance. It is necessary to add that definite conclusions cannot be drawn from our results due to the relatively small number of subjects. During our analysis, considerable individual differences were shown in both subject groups.

Further steps of the research include increasing the sample size and introducing male subjects, which would provide an opportunity to uncover potential gender differences. Another addition is the analysis of prolongations, which will also be undertaken in the future. The authors hope that the results of this research can broaden the understanding of the speech production of mildly intellectually disabled adults and aid the development of therapy methods.

Acknowledgements

We thank the Department of Phonetics, as well as the MTA–ELTE "Lendület" Lingual Articulation Research Group at ELTE for the location and tools supplied for the current research.

References

Boersma, P. & D. Weenink. 2018. *Praat: Doing phonetics by computer* (version 6.0.43). http://www.praat.org/ (accessed 1 January 2019).

Bóna, J. 2018. Disfluencies and disfluency clusters in cluttered, stuttered and typical speech. *Beszédkutatás* 26: 221–235.

Bóna, J., A. Imre, A. Markó, V. Váradi & M. Gósy. 2014. GABI – Gyermeknyelvi beszédAdatBázis és Információtár [GABI - Children's Speech and Informational Database]. *Beszédkutatás* 22: 246–251.

Coppens-Hofman, M. C., H. R. Terband, B. A. Maassen, H. M. Van Schrojenstein Lantman-De Valk, Y. Van-Zaalen-op't Hof & A. F. M. Snik. 2013. Disfluencies in the speech of adults with intellectual disabilities and reported speech difficulties. *Journal of Communication Disorders* 46(5–6): 484–494. https://doi.org/10.1016/j.jcomdis.2013.08.001

ICD-11 = International Classification of Diseases. 11th Revision. The global standard for diagnostic health information. 2018. https://icd.who.int/en/ (accessed 29 May 2019).

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

Lukács Á. & B. Kas. 2014. Nyelvelsajátítás és értelmi fogyatékosság [Language acquisition and mental disability]. In: Pléh Cs. & Á. Lukács (eds.): *Pszicholingvisztika 1–2.: Magyar pszicholingvisztikai kézikönyv* [Psycholinguistics 1–2. Hungarian handbook on psycholinguistics] Budapest: Akadémiai Kiadó, 1383–1404.

Roberts, P. M., A. Meltzer & J. Wilding. 2009. Disfluencies in non-stuttering adults across sample lengths and topics. *Journal of Communication Disorders* 42(6): 414–427. https://doi.org/10.1016/j.jcomdis.2009.06.001

Rossi, N. F., A. Sampaio, Ó. F. Gonçalves & C. M. Giacheti. 2011. Analysis of speech fluency in Williams syndrome. *Research in Developmental Disabilities* 32(6): 2957–2962. https://doi.org/10.1016/j.ridd.2011.05.006

Tóth, A. 2017. *A spontán beszéd a nem és az életkor függvényében gyermek- és fiatal felnőttkorban* [Spontaneous speech in relation to gender and age in childhood and adolescence]. Ph.D. dissertation. Eötvös Loránd University, Budapest, Hungary.

Special day on (dis)fluency in children's speech
Preface

DiSS 2019 featured a co-located event on September 14th, 2019 which focused on (dis)fluency in children's speech. Following are the abstracts from the various presentations given that day. Full paper versions of the research presented is expected to be published in 2020. The abstracts are provided here as a conveniece to readers and as a record of the DiSS 2019 event. However, researchers are encouraged to read and cite the full versions of the following papers once they appear.

Implications of a developmental approach for understanding spoken language production

Melissa A. Redford
Department of Linguistics, University of Oregon, Eugene, Oregon, USA

Current approaches to spoken language production focus on adult behavior and so make assumptions that, when taken to their logical conclusion, are inconsistent with spoken language acquisition. This inconsistency represents a significant weakness for production theories given that adults were children once. If childhood shapes adulthood through an incremental development process where new structures build on existing structures, then we would expect speech and language acquisition to shape the representations that allow adults to turn language into the action that is speech. This is the perspective I have taken in proposing a developmentally sensitive theory of production (see Redford, 2015; Redford, in press; Davis & Redford, in press); one where linguistic representation is understood to evolve continually with lifespan changes in language ability, executive functioning (i.e. cognition pertaining to sequencing and planning), and speech motor control. A practical implication of the theory is a novel framework for the investigation of normal and disordered spoken language production. The plan for this talk is to illustrate how the theory has shaped my own empirical work on child and adult speech and language. More specifically, my plan is to elaborate on the central premise that speech motor skill development interacts with language acquisition to define the hierarchical structure of language, including temporal patterns that have been identified as relevant to the perception of spoken language rhythm. Phrasing will serve as a first example of this interaction: work on children's pausing patterns and adults' speech error patterns coupled with widely-documented patterns of child language are used to suggest that internal prosodic phrase boundaries, like layers in sedimentary rock, define eras; in the case of prosodic boundaries, these eras represent utterance lengths and predominant chunking patterns across developmental time. Thinking about phrasing in this way leads us to ask interesting new questions about the coordination of language planning and breathing. So, for example, we are currently investigating anticipatory and recovery effects on breath intakes with the long-term goal of better understanding the process by which children learn to breathe at grammatical junctures. Syllable reduction in children's speech will serve as a second example of the interaction between speech motor skill development and language acquisition: a review of work on rate and long-distance coarticulation suggest that it is the acquisition of articulatory timing skills, not metrical structure or chunking per se, that conditions the development of grammatical word reduction and the emergence of adult-like English rhythm in children's speech. Understanding English speech rhythm acquisition in this way requires a theory of phonological representation that incorporates relative timing information. I will speculate on how adult listener expectations might encourage the development of such representations with reference to new data. Overall, the talk will outline a wide-ranging research program structured by a developmental approach to spoken language production. Important conceptual challenges not yet addressed and other gaps in the program will be identified as objects of future research.

References

Davis, M. & M. A. Redford. In press. The emergence of perceptual-motor units in a production model that assumes holistic speech plans. *Frontiers in Psychology*.

Redford, M. A. 2015. Unifying speech and language in a developmentally sensitive model of production. *Journal of Phonetics*, 53: 141–152. https://doi.org/10.1016/j.wocn.2015.06.006

Redford, M. A. In press. Speech production from a developmental perspective. *Journal of Speech, Language, and Hearing Research*. https://doi.org/10.1044/2019_JSLHR-S-CSMC7-18-0130

The role of disfluencies in language acquisition and development of syntactic complexity in children

Ivana Didirkova [1], Christelle Dodane [2] and Sascha Diwersy [2]
[1]University Paris 8, EA1569 TransCrit – LeCSeL & University Paris 3, UMR7018 Laboratoire de Phonétique et phonologie, CNRS, France
[2]University Paul-Valéry Montpellier 3, UMR5267 Praxiling, CNRS, France

The language acquisition process is, among others, characterized by a period during which children produce an important number of disfluencies. This period roughly corresponds to ages 2 to 3 and is known as being transient. It coincides with the beginning of the production of longer utterances with a higher syntactic complexity, requiring more complex motor planning (Peters et al., 1989). These increased demands may be too heavy for the child to process, leading to an increased amount of speech disruptions. Thus, it is considered that the study of disfluencies would reveal a lot about the difficulties encountered by children when constructing their utterances (Yaruss et al., 1999). We then can presume that disfluencies will be more frequent with the increasing syntactic complexity.

More specifically, the aim of this study is to investigate the influence of syntactic complexity on the number of disfluencies. We made the choice to use longitudinal data containing speech productions of 4 children (2 boys and 2 girls) aged 1 to 6 (Morgenstern & Parisse, 2012), registered monthly at home, in spontaneous interaction with their parents. These productions were entirely transcribed according to the CHAT standards (CHILDES system, MacWhinney, 2000) within the CLAN[1] software. Disfluencies were annotated directly in the transcriptions. We then extracted automatically the total number of disfluencies per utterance (henceforth TND). Our analyses focus on 56400 utterances (14100 utterances per child; $sd = 2211$ utterances). To estimate the degree of syntactic complexity, we calculated the Mean Length of Utterances (MLU, Brown, 1973) and the mean number of verbs per utterance for each child at each age. A Pearson's correlation coefficient showed a strong linear relation between those measures (Figure 1; test statistics: $cor = 0.77$, $t = 12.298$, $df = 105$, $p < 0.005$).

We also found a strong correlation between MLU and the TND ($cor = 0.65$, $t = 8.796$, $df = 105$, $p < 0.005$). Overall, an increase in the number of disfluencies is observed as the age increases. More precisely, before age 2, only a few disfluencies are found in the speech production of our four participants. This small amount can be explained by the nature of produced utterances (absence of function words, often replaced by fillers). Between ages 2 and 4, an important variation is observed, accompanied by a constant increase of the number of disfluencies. Finally, after 4 years, and except some of the recording sessions (outliers in the Figure 2), children tend to produce shorter utterances with less disfluencies, which would correspond to a slow-down phase in the syntax acquisition.

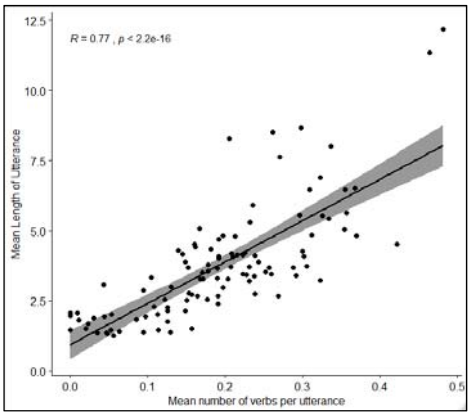


Figure 1: MLU and mean verb per utterance correlation.

These results will be detailed for all participants and will be linked with different periods in language acquisition. Furthermore, we will present other measures of syntactic complexity during the DiSS 2019 conference, such as the *Index of Syntactic Complexity* (Szmrecsanyi, 2004), the *Developmental Sentence Scoring* (DSS, Lee, 1966) used both for the evaluation of syntactic performance in children and for the diagnostic of language disorders (Rheinhardt, 1972) and the *Index of Productive Syntax* (IPSyn, Moyle & Long, 2013) allowing to evaluate and to quantify the syntactic complexity in young children.

Acknowledgment

This research was supported by the BENEPHIDIRE project (PI: Fabrice Hirsch, ANR-18-CE36-0008).

Notes

[1] Data used in this study are part of the COLAJE project (Morgenstern & Parisse, 2012). They are available on the website of the project on the website of the CHILDES project (MacWhinney, 2000).

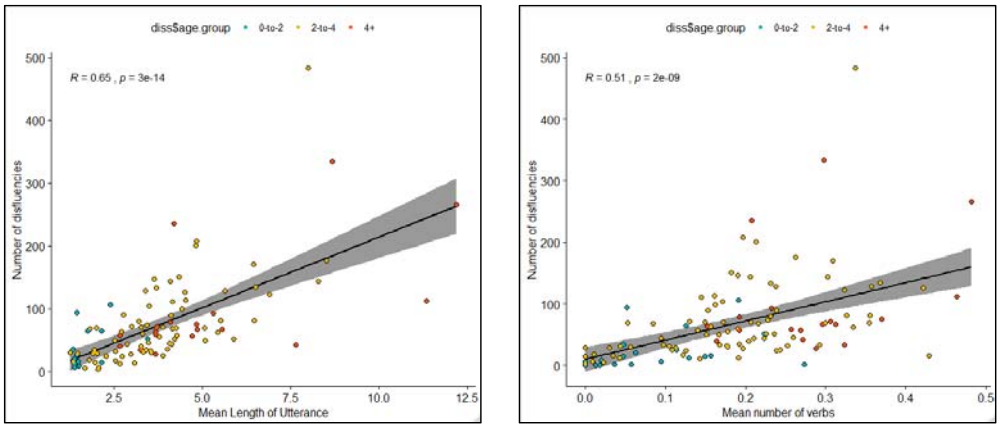


Figure 2. Correlation between MLU (0.65), mean number of verbs per utterance (0.51, t = 6.501, df = 119, p < 0.005) and TND.

References

Brown, R. 1973. *A first language: the early stages*. Cambridge, MA: Harvard University press. https://doi.org/10.4159/harvard.9780674732469

Lee, L. L. 1966. Developmental sentence types: a method for comparing normal and deviant syntactic development. *The Journal of Speech and Hearing Disorders* 31(4): 311-330. https://doi.org/10.1044/jshd.3104.311

MacWhinney, B. 2000. *The CHILDES Project: Tools for analyzing talk*, 3rd edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Morgenstern, A. & C. Parisse. 2012. The Paris Corpus. *French language studies* 22(1): 7–12. https://doi.org/10.1017/S095926951100055X

Moyle, D. M. & S. Long. 2013. Index of Productive Syntax (IPSyn). In: F. R. Volkmar (ed.), *Encyclopedia of Autism Spectrum Disorders* 1566-1568. New York: Springer.

Peters, H. F. M., W. Hulstijn & C. W. Starkweather. 1989. Acoustic and physiological reaction times of stutterers and nonstutterers. *Journal of speech and hearing research* 32(3): 668–680. https://doi.org/10.1044/jshr.3203.668

Rheinhardt, K. 1972. The Developmental Sentence Scoring Procedure. In: *Independent Studies and Capstones*, Paper 314. Program in Audiology and Communication Sciences, Washington University School of Medicine.

Szmrecsanyi, B. 2004. On operationalizing syntactic complexity. In: C. F. Grard Purnelle & A. Dister (eds.), *7th International Conference on the Statistical Analysis of Textual Data* (JADT 2004), 10–12 March 2004, Louvain-la-Neuve, Belgium, vol. 2, 1032–1039.

Yaruss, J. S., R. M. Newman & T. Flora. 1999. Language and disfluency in nonstuttering children's conversational speech. *Journal of Fluency Disorders* 24(3): 185–207. https://doi.org/10.1016/S0094-730X(99)00009-1

Filled pauses in children's spontaneous speech – aspects from timing and complexity

Viktória Horváth and Valéria Krepsz
Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary

Filled pauses (FP) can have many functions: indicates the speaker's current speech planning problem, provides time for self-monitoring and correction processes, and can play a role in organizing turn-takings in conversations. Filled pause is a universal phenomenon, but its realizations are language-specific. The realizations of FPs depend on many factors, e.g. the type of speech, the characteristics of the speaker, as well as the speakers' age and the grammatical structure of the utterance. Previous results examined the interaction of speech rate, length, and grammatical complexity, found that clausal constituents were highly correlated with MLU (mean length of utterances). In addition, other not significant trends were detectable, for example the longer length and lower rate attended more disfluent utterances (Sawyer et al., 2008). Connection between (i) the utterance length and complexity, (ii) as well as the length of the utterances (i.e. MLU) and the frequency of the child's speech disfluency were confirmed by another study as well (Zackheim & Conture, 2003).

However, it is questionable how these tendencies develop in childhood in another type of (agglutinating) language. Because of the complex morphological system (included big amount of irregular form of nouns and verbs, suffixes and suppletive forms) the acquisition of the suffix system can be extended until early school-age (Pléh, 2006). Therefore, this study examined relationships among utterance length, speech rate, syntactic complexity, and children's age in spontaneous speech. According to the hypotheses of this research, the grammatical complexity will gradually increase with age, but with the incidence of filled pause the correlation will be increasingly weaker.

Participants were 10 four-year-old, 10 six-year-old and 10 eight-year-old children (5 boys and 5 girls who were selected from the GABI database (Bóna et al., 2014). The frequency, the phonetic form, the position and the duration of FPs were analyzed on the one hand. On the other hand, the number of clauses, the mean length of utterance (MLU) in morphemes, and articulation rate, measured in syllables per second were analyzed from the children's conversational speech. In addition, the grammatical complexity (DSS, Developmental Sentence Scoring) of the utterances was scored in a point system that was built on the peculiarities of the Hungarian grammatical and lexical system (Lee & Carter, 1971, adopted to Hungarian Gerebenné Várbíró et al., 1992). The interactions between the occurrence and the duration of FPs, utterance length, articulation rate, and grammatical complexity were measured and have been compared by the age of the speakers.

Results show that there was significant difference in the frequency of FPs depending on children's age. However pairwise comparisons showed statistical difference only between 6- and 8-year-olds. The most frequent type of FP was schwa, irrespective of children's age. The statistical analyses revealed that the children's age and the position of schwa affected FP's duration. The tendencies were similar both in the FP's frequency and in the duration of schwas between 4- and 8-year-olds: the values decreased between 4 and 6, then increased between 6 and 8 years (Figure 1). The mean value and the standard deviation of DSS gradually increased with age. Besides, the individual differences were great at all three ages. Grammatical complexity has correlated with the realization of FPs. Therefore, we can

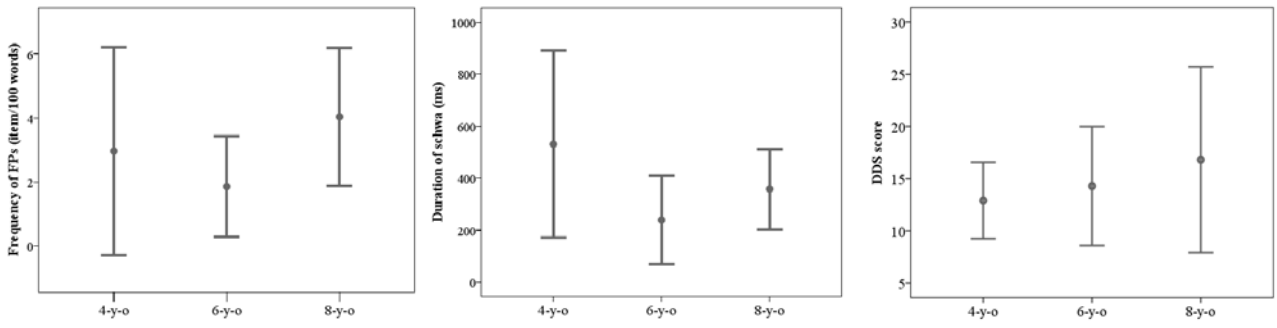


Figure 1: The frequency of FPs (left), the duration of schwa-like FPs (middle) and the DSS score (right) in the three age groups' spontaneous speech

conclude, that children use different strategies at different ages to resolve the disfluencies of spontaneous speech, that appears in the production patterns of spontaneous speech.

Acknowledgement

This research was supported by the Hungarian National Research, Development and Innovation Office of Hungary, project No. K-120234.

References

Bóna, J., A. Imre, A. Markó, V. Váradi & M. Gósy. 2014. GABI – Gyermeknyelvi beszédAdatBázis és Információtár [GABI – Children's Speech and Informational Database]. *Beszédkutatás* 22: 246–251.

Gerebenné Várbíró, K., M. Gósy & M. Laczkó. 1992. *Spontán beszédmegnyilvánulások szintaktikai elemzése DSS technika segítségével* [Syntactic analysis of spontaneous speech using DSS technique]. Kézirat. Budapest.

Pléh, Cs. 2006. A gyermeknyelv [The language of children]. In: Kiefer, F. (ed.), *Magyar nyelv* [Hungarian language], 753–782. Budapest: Akadémiai Kiadó.

Sawyer, J., H.-C., Chon & N. G. Ambrose. 2008. Influences of rate, length, and complexity on speech disfluency in a single-speech sample in preschool children who stutter. *Journal of Fluency Disorders* 33(3): 220–240. https://doi.org/10.1016/j.jfludis.2008.06.003

Zackheim, C. T. & E. G. Conture. 2003. Childhood stuttering and speech disfluencies in relation to children's mean length of utterance: a preliminary study. *Journal of Fluency Disorders* 28(2): 115–142. https://doi.org/10.1016/S0094-730X(03)00007-X

Filler words in children's and adults' spontaneous speech

Mária Gósy
Research Institute for Linguistics, Hungarian Academy of Sciences, and
Department of Phonetics, ELTE University, Budapest, Hungary

Filler words and filled pauses as hesitation phenomena are connected with prospective and retrospective tasks of speech production (e.g. Shriberg, 2001; Clark & Fox Tree, 2002; Watanabe et al., 2008). For definitions, (i) a filler word can be any word or phrase that is inserted into the main message of the utterance with a function that is associated with conceptual and other difficulties of speech planning; (ii) a filled pause is a meaningless sound sequence containing a neutral vowel of various durations (in our case) with a function of revealing speech planning and execution problems. Both filler words and filled pauses have language-specific nature. In children with typical language development, the occurrence of filler words and filled pauses increase with age, linguistic complexity, longer communication units and larger amount of different words produced (e.g. Thordardottir & Weismer, 2002; Neuberger & Gósy, 2014). In the beginning, children are assumed to simply imitate the filler words and filled pauses, later on they acquire their functions and the strategy of their use (e.g. Schiro, 2003; Furman & Özyürek, 2007). Two questions arise whether (i) there is a specific interrelation of filler words and filled pauses during language acquisition, and (ii) there is any specific function that differentiates the use of filler words and filled pauses across ages? We hypothesized that children use less filler words than filled pauses while young adults use more filler words than filled pauses. We assumed that the occurrence and position of these hesitation phenomena show a developing strategy to indicate speech planning differences to addressees.

Spontaneous speech samples of 42 speakers (7-year-old children, 10-year-old children and young adults; half of them were females in all groups) selected from two large databases (GABI and BEA) were analyzed. All speakers were monolingual native speakers of Hungarian. No hearing or speaking disability was reported among them. The topics of the adults' narratives concerned their families, work, and hobbies while children spoke about family, school, holiday and hobbies. Each filled pause and filler word was identified and manually annotated using Praat software. Their placement according to the context was identified using the criteria whether they occurred (i) within a phrase, (ii) at phrase boundaries or, (iii) at the very beginning of a phrase. Close to 2,600 items were produced by the participants.

Results indicated an increasing number of both filled pauses and filler words across ages. The youngest children had more filled pauses than filler words. Ten-year-olds produced similar amount of filled pauses and filler words. Young adults' speech samples contained more filler words than filled pauses. Filled pauses were more frequent at phrase boundaries than within phrases with 10-year-olds while the opposite could be noticed with young adults. No difference was found with 7-year-olds. Filler words, however, occurred more frequently at phrase boundaries than within phrases with 7-year-olds and adults. They were less frequent at phrase boundaries with 10-year-olds. Both analyzed hesitation phenomena occurred rarely at the very beginning of the phrases in all groups.

Both the occurrences and positions of filler words and filled pauses show an emerging strategy reflecting a communicative signal to cognitive, lexical and execution difficulties. The differences depending on age are discussed in terms of various strategies children and young adults use.

Acknowledgement

This research was supported by the Hungarian National Research, Development and Innovation Office of Hungary, project No. K-120234.

References

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

Furman, R. & A. Özyürek. 2007. Development of interactional discourse markers: Insights from Turkish children's and adults' oral narratives. *Journal of Pragmatics* 39(10): 1742–1757. https://doi.org/10.1016/j.pragma.2007.01.008

Neuberger, T. & M. Gósy. 2014. A cross-sectional study of disfluency characteristics in children's spontaneous speech. *Govor* 31(1): 3–28.

Schiro, M. 2003. Genre and evaluation in narrative development. *Journal of Child Language* 30(1): 165–195. https://doi.org/10.1017/S0305000902005500

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

Thordardottir, E. T. & S. E. Weismer. 2002. Content mazes and filled pauses on narrative language samples of children with specific language impairment. *Brain and Cognition* 48(2–3): 587–592. https://doi.org/10.1006/brcg.2001.1422

Watanabe, M., K. Hirose, Y. Den & N. Minematsu. 2008. Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication* 50(2): 81–94. https://doi.org/10.1016/j.specom.2007.06.002

Use of words in story-telling data of Chinese-speaking hearing and hearing-impaired children

Yi-Fen Liu [1] and Shu-Chuan Tseng [2]
[1]Department of Information Engineering and Computer Science, Feng-Chia University, Taichung, Taiwan
[2]Institute of Linguistics, Academia Sinica, Taipei, Taiwan

Purpose

While it is likely to mitigate the discrepancy of spoken language performance between hearing and hearing-impaired children with intervention and verbal training, hearing impairment may still affect the use of spoken language in different aspects. This study aims to observe how spoken words are produced in story-telling speech data by three subject groups (NH, CI, HA), including word category, IPU length, duration, and reduction types.

Subjects

79 preschool children with normal hearing (NH) participated in the recording, aged 2;11–6;3 (*median* = 4;11). Thirty-three children with hearing impairment were recorded, aged 3;3–6;11 (*median* = 5;1). Among them, 22 children wore traditional hearing aids (HA, moderate-severe), and 11 children were fitted with cochlear implants (CI, profound).

Data

The children were instructed to tell the story of the Hare and Tortoise presented with six picture cards illustrating the story content in a fixed order. HA and CI children were recorded during their regular AVT session in sound-proof classrooms (Dornan et al., 2007). NH children were either recorded at Academia Sinica in sound-proof studios or in quiet classrooms at their kindergarten. The data were digitized at a sampling rate of 44.1 kHz with a 16-bit quantization. The recorded content was transcribed in traditional Chinese characters including annotations of pauses and paralinguistic sounds. An automatic phone aligner developed at Academia Sinica was used to obtain timestamps for syllable and phoneme boundaries (Liu et al., 2014). Word boundary information was generated by integrating results of the CKIP automatic word segmentation and POS tagging system (Chen et al., 1996).

Methodology

In order to examine word category from a more general perspective, we transformed all CKIP POS tags to Universal POS categories (Universal Dependencies webpage). For word-level reduction, we adopted the derivation approach proposed by Liu et al. (2016) by comparing surface phone sequences generated from free phone recognition results with the phoneme sequences directly derived from the standard phonological forms of the disyllabic words. Four reduction degrees are differentiated from minor to severe: CAN < MSD < NUM < SYM. CAN (canonical-form-like) is produced with no omission of cross-boundary consonants, while MSD omits at least one of the cross-boundary consonants. NUM loses all cross-boundary consonants, while the two nuclei are clearly present, yet with a vague syllable boundary. SYM (syllable merger) merges two syllables into one. Figure 1 shows examples of these four reduction types of the word *ran2hou4* (then) produced by a child with normal hearing.

Results

As a result, patterns in terms of word category are quite similar across three subject groups, while individual groups show slight differences in word use. NH children used more prepositions and

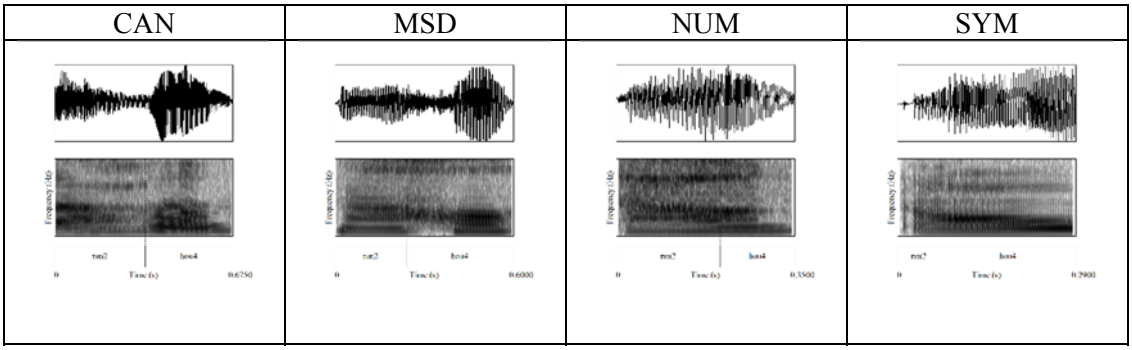


Figure 1. Four disyllabic word reduction types in child speech.

pronouns, whereas CI and HA children used more nouns and verbs (Figure 2). In terms of IPU (inter-pausal units), all three subject groups also show similar tendencies, except for a clear smaller number of single word IPU in NH children (Figure 3).

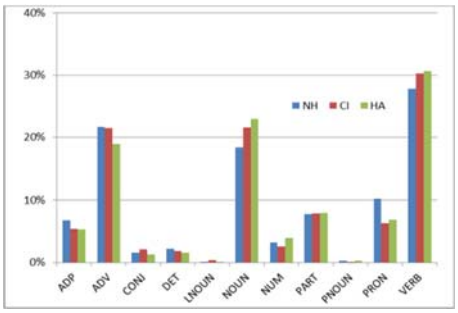


Figure 2. Word categories

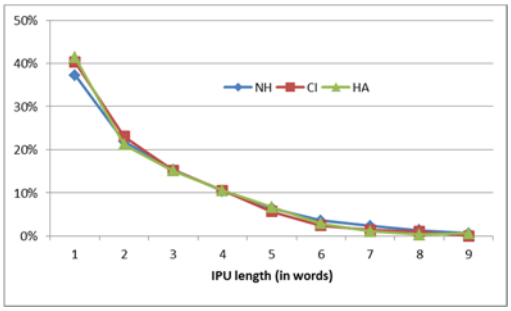


Figure 3. IPU length in words

Concerning prosodic properties, duration patterns are examined. In Figure 4, NH, CI, and HA children show rather similar word duration patterns (in terms of mean, but we are currently doing more detailed analysis). Please note that both CI and HA children received AVT training and the duration of wearing hearing aids differ greatly. However, in spite of similar duration patterns, NH children clearly differentiate themselves from CI and HA children in the analysis of word-level reduction types. The general pattern is similar to that reported in Liu et al. (2016) for adults' speech, where CAN and SYM are preferred than MSD and NUM. NH children apparently produced more CAN and less SYM than both of the CI and HA children (Figure 5). For adults' speech, the preference for SYM in high-frequency words is supportive of possible phonetic representation forms in the mental lexicon. But for preschool children's speech, the only conclusion we can draw at this stage is that the speech of CI and HA children seems to be much more reduced than that of NH children on the basis of information extracted from the speech signal. Beyond the preliminary results presented here, further studies on the phonetic representations of spoken words produced by these three subject groups are currently in process.

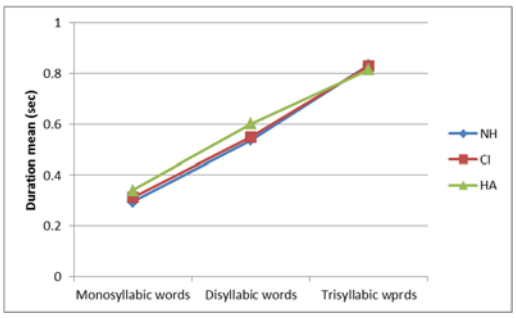


Figure 4. Word duration mean

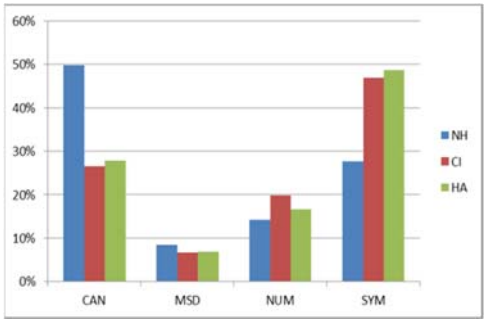


Figure 5: Disyllabic word reduction types

References

Chen, K.-J., C.-R. Huang, L.-P. Chang & H.-L. Hsu. 1996. SINICA CORPUS: Design Methodology for Balanced Corpora. In: B.-S. Park & J.-B. Kim (eds.), *Language, Information and Computation: Selected Papers from the 11th Conference on Language, Information and Computation* (PACLIC 11), 20–22 December 1996, Seoul, Korea, 167–176.

Dornan, D., L. Hickson, B. Murdoch & T. Houston. 2007. Outcomes of an auditory-verbal program for children with hearing loss: A comparative study with a matched group of children with normal hearing. *Volta Review* 107(1): 37–54.

Liu, Y.-F., S.-C. Tseng & R. Jang. 2014. Phone boundary annotation in conversational speech. In: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of LREC 2014*, 26 May 2014, Rejkjavik, Iceland, 848–853.

Liu, Y.-F., S.-C. Tseng, and R. Jang. 2016. Deriving disyllabic word variants from a Chinese conversational speech corpus. *Journal of the Acoustical Society of America* 140(1): 308–321. https://doi.org/10.1121/1.4954745

Patterns of lingual CV coarticulation in Hungarian children's speech: The case of stops

Alexandra Markó [1,2], Tamás Gábor Csapó [2,3], Márton Bartók [1,2], Tekla Etelka Gráczi [2,4], and Andrea Deme [1,2]
[1]Department of Applied Linguistics and Phonetics, Eötvös Loránd University (ELTE), Budapest, Hungary
[2]MTA–ELTE "Lendület" Lingual Articulation Research Group, Budapest, Hungary
[3]Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Budapest, Hungary
[4]Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary

Coarticulation is a complex mechanism, which involves multiple articulators' finely coordinated actions over time in order to produce fluent speech. Fluent speech is the result of coarticulation.

Coarticulation is segment dependent in the sense that some of the speech sounds are more resistant, while another group of the speech sounds is more easily affected by coarticulation. Furthermore, some segments tend to affect the adjacent segments to a larger extent than others. For instance, consonants with lingual places of articulation might have an effect on the adjacent vowels' articulation, while bilabial consonants are less likely to do so (see e.g. the DAC model by Recasens et al., 1997).

With regard to coarticulation, children's speech has been the subject of numerous studies, from the very early stages of language acquisition, in several languages (see an overview by Mildner, 2018). These studies analyze both the articulatory and the acoustic domain of speech. However, Hungarian is rather understudied in this respect; coarticulatory mechanisms of children's speech have hardly been examined so far. The present study aims to reveal some patterns of CV coarticulation at various ages using the method of ultrasound tongue imaging and formant analysis.

In an earlier study, trisyllabic, nonsense $V_1C_1V_1C_1V_1$ structured sequences (with 9 vowels and 9 consonants of Hungarian, in all combinations) were recorded, which were produced by 6 children and one adult (Table 1, Markó et al., in press), in two repetitions.

Table 1: The age (years;months) and gender of the participants

ID	0161	0155	0163	0160	0162	0154	0048
Gender	F	M	F	M	F	M	F
Age	7;10	10;0	11;1	11;5	12;10	14;8	43;5

The speakers' tongue movement was recorded in midsagittal orientation using the "Micro" ultrasound system (Articulate Instruments Ltd.) with a 2–4 MHz / 64 element 20 mm radius convex ultrasound transducer at 83 fps. The participants' head was fixed relative to the transducer with a headset. The speech signal was recorded with an omnidirectional condenser microphone at 44.1 kHz sampling rate. In the earlier study, the analysis focused on the variability of vowel articulation. Tongue contours were manually traced in the ultrasound images obtained at the temporal midpoint of the vowels, in two repetitions. The coarticulatory effect of the adjacent consonants on the vowels was analysed and compared with respect to places of articulation (we analysed labial and lingual consonants), using the nearest neighbour distance (NND, Zharkova et al., 2011) method. However, the results were inconclusive in part, and did not meet the expectations that lingual places of articulation of the neighbouring consonants have a larger effect on the vowels than the labial ones. Therefore, in the present study, a different method is planned to be applied for the reanalysis of the same dataset.

Sequences of four vowels and four consonants were selected from the corpus (Table 2). Vowels from the second syllable (which is unstressed due to the fixed first syllable stress of Hungarian) of the sequences were analysed both in the articulatory and the acoustic domain.

Table 2: Sequences selected for analysis (target vowels are indicated with bold)

	/i/	/u/	/ɛ/	/ɒ/
/p/	/ip**i**pi/	/up**u**pu/	/ɛp**ɛ**pɛ/	/ɒp**ɒ**pɒ/
/t/	/it**i**ti/	/ut**u**tu/	/ɛt**ɛ**tɛ/	/ɒt**ɒ**tɒ/
/c/	/ic**i**ci/	/uc**u**cu/	/ɛc**ɛ**cɛ/	/ɒc**ɒ**cɒ/
/k/	/ik**i**ki/	/uk**u**ku/	/ɛk**ɛ**kɛ/	/ɒk**ɒ**kɒ/

We trace tongue contours manually at five time points (0%, 25%, 50%, 75% and 100% of the total vowel duration) of the selected vowels. At the same five time points, we also estimate first and second formants' values. Finally, we also extract palate contours from the ultrasound images which were recorded during wet swallowing (Epstein & Stone, 2005).

In the articulatory domain, we plan to perform two comparisons. First, we calculate the closest distances between the palate contour and the tongue contours in each vowel's case for each measurement point using the NND method, and then we average these values for the two repetitions and the five time points of the vowel at hand. We analyse this measure as a function of the adjacent consonants' place of articulation. The tongue contours of the vowels in the same contexts are also planned to be averaged for a qualitative analysis.

In the acoustic analysis, we average the formant values of the vowels in the same contexts and compare them with respect to the place of articulation of the adjacent consonant.

In accordance with previous findings, we expect larger CV effects in the lingual places of articulation than in the bilabial one. However, different patterns are expected as a function of the age of the participants, as well as the articulatory features of the analysed vowels.

References

Epstein, M. A. & M. Stone. 2005. The tongue stops here: Ultrasound imaging of the palate. *Journal of the Acoustical Society of America* 118(4): 2128–2131. https://doi.org/10.1121/1.2031977

Markó, A., T. G. Csapó, A. Deme, T. E. Gráczi & M. Bartók. In press. Gyermekek lingvális artikulációjának variabilitása magánhangzós nyelvkontúrok alapján [Variability of children's lingual articulation based on vowel tongue contours].

Mildner, V. 2018. Aspects of coarticulation. In: M. Gósy & T. E. Gráczi (eds), *Challenges in analysis and processing of spontaneous speech*, 27–48. Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences.

Recasens, D., M. D. Pallarès & J. Fontdevila. 1997. A model of lingual coarticulation based on articulatory constraints. *Journal of Acoustical Society of America* 102(1): 544–561. https://doi.org/10.1121/1.419727

Zharkova, N., N. Hewlett & W. J. Hardcastle. 2011. Coarticulation as an indicator of speech motor control development in children: An ultrasound study. *Motor Control* 15(1): 118–140. https://doi.org/10.1123/mcj.15.1.118

Characteristics of disfluencies in teenagers' spontaneous speech and topic based narratives created during the lessons

Mária Laczkó
Faculty of Pedagogy, Kaposvár University, Kaposvár, Hungary

Introduction

Our continuous speech is not fluent as the pronunciation of speech sounds and silent periods alternate with each nother and various types of interruptions may occur in the process. In Levelt's theory (1989, 1993) the process of spontaneous speech consists of the plan of thoughts and linguistic form and the articulation process. As the speaker can hesitate about the topic or the linguistic forms, and some mistakes can also occur in the articulation process, the spontaneous speech is not fluent and interrupted by various kind of disharmonic phenomena (Gósy, 2002, Postma et al., 1990, Shriberg, 2001). Hungarian scholars categorized them as the types of uncertainties like repetition, restart, filled pauses, fillers and the mistakes like false start, mistakes of order (methatezis, anticipation, perseveration), slips of the tongues, or grammatical mistakes.

These disfluencies were analysed mainly in adults' speech and there are less data regarding the analysis of them in child language, especially among teenagers. The scholars have analysed the interrelation between the complexity of sentences and the frequency of occurrences of various types of disfluencies (Silverman & Bernstein Ratner, 1997). In one of the Hungarian work the influence of the type of the text on the frequency of interruptions was analysed (Laczkó, 2010). There are also some experimental data in terms of the ratio of disfluencies occurring in dialogues (Libárdi, 2015), or in the spontaneous narratives among the students who are preparing for their final exam (Vallent, 2009).

However, there are some data concerning the distribution of various types of disfluences in teenagers' spontaneous speech there is no experimental data about them in the special communication situation when they create the narratives during the lessons as the responses based on the given topic.

This communication situation is different from the spontaneous speech, consequently it demands other kind of planning processes. However, the topic based narrative during the lesson is more common for the students it might be more difficult which can be followed both on the segmental and suprasegmental levels. Our previous research had already proved this hypothesis as the type/token index and the KFM score of the sentences in the two types of speeches were different from each other. There was also significant difference in terms of the speech tempo and articulation tempo categories. The ratio of the pauses even the duration of them was also different in the two kind of speeches. (Laczkó, in press).

On the basis of these findings the actual research question is to analyse the interruptions in the two communication situations. The aim of the presentation is to define the characteristic features of different types of disfluencies occurring in their spontaneous speech versus the topic based narratives created during the lessons. We answer the question whether the difficulty of the planning process how can occur in the ratio of disharmonic phenomena called uncertainties and mistakes. We also answer the question whether what types of disfluencies belonging to the mentioned categories can occur in the topic based narrative versus the spontaneous speech and to what extent.

Our previous hypothesis was that the ratio of uncertainties and mistakes in the topic based narrative is different from the ratio occurring in spontaneous speech. The difference can be followed especially in terms of the types of uncertainties, so the pattern of these kind of disfluencies is showing different pattern from the spontaneous speech. We also thought that the pattern of mistakes can also be different in the two communication situations.

Method, material

In order to answer the questions and to prove the hypothesis the experiment was carried out with the participation of teenagers. Their spontaneous speech and topic based narratives created during the lesson were digitally recorded. The same students participated in both communication situations. The students were selected from 15-year-old and 17-year-old secondary technical students (with normal hearing and intelligence), the number of them was 5-5 in each situation, the all number of students's speech samples was 20, the average duration of the speech samples was more than 2 minutes.

From the speech samples the interruptions and disfluencies were gathered and they were categorized in different types. For the categorization the definition of previous categories (cf. Gósy, 2009,

2015) were used. In the analyses the silent pauses were not taken into account as the interruption (cf. Fox Tree, 1995). We also analysed the frequency of the types of disfluences (it is given in the number of disfluencies per 100 word, and in terms of the number of them per minute). Among the types of interruptions, the most frequent categories (filled pauses, prolongation, pause in the word) were also analysed in terms of their temporal characteristics and function. For the acoustic analysis the Praat program (Boersma & Weenink, 2008) was used, while the statistical analysis was done by the SPSS 13.00 version.

Results

The result have confirmed all the hypothesis. Both the frequency of interruptions, even the main categories were different in the two communication situations, as the topic based narratives created during the lessons contained much more disfluencies independently the age, and the difference was significant. The ratio of various types of uncertainties were also different in the two kind of speeches in both of the age groups, so the pattern of uncertainties was influenced by the type of the text. The topic based narratives contained more mistakes too, and the distribution of them was again different from the distribution of mistakes found in the spontaneous speech. The main important disfluencies in the topic based narratives beside the filled pauses and prolongations was the category of pause in the word, versus the spontaneous speech where the fillers. Among the mistakes the false word activation was the main important disfluency type regarding the topic based narrative versus the spontaneous where the perseveration. We also found the difference in between the speeches created in the two kind of communication situations in terms of temporal characteristic of the disfluences and regarding their functions as well.

As the examination is focused on the analysis of the Hungarian students' topic based narratives created during the lessons for the first time the presentation emphasizes the detailed results and the consequences both in linguistic and pedagogical aspects.

References

Boersma, P. & D. Weenink. 2008. Praat: Doing phonetics by computer (version 5.0.1). http://www.praat.org/ (accessed 10 September 2015).

Fox Tree, J. E. 1995. The effect of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language* 34(6): 709–734. https://doi.org/10.1006/jmla.1995.1032

Gósy, M. 2002. A megakadásjelenségek eredete a beszédprodukció tervezési folyamatában [The origin of interruption phenomena in the speech production planning process]. *Magyar Nyelvőr* 124: 192–204.

Gósy, M. 2009. Önjavítási stratégiák a beszédben gyermekeknél és felnőtteknél [Self-improvement strategies in speech in children and adults]. In: B. Vilmos (ed.), Quo vadis philologia temporum nostrorum. Korunk civilizációjának nyelvi képe [Quo vadis philologia temporum nostrorum. The linguistic picture of the modern civilization], 141–150. Budapest: Tinta Könyvkiadó, Budapest.

Gósy, M. (ed.). 2015. Diszharmóniás jelenségek a beszédben [Disharmonic phenomena in speech]. Budapest: MTA Nyelvtudományi Intézet.

Levelt, W. J. M. 1989. *Speaking. From intention to articulation,* Cambridge, MA, USA: A Bradford Book.

Levelt, W. J. M. 1993. Accessing words in speech production. Stages, processes and representations. In: Levelt, W. J. M. (ed.), *Lexical Access in Speech Production*, 1–22. Cambridge, MA, USA: Blackwell. https://doi.org/10.1016/0010-0277(92)90038-J

Laczkó, M. 2010. Megakadások a spontán és a szónoki beszédben [Interruptions in spontaneous and rhetorical speech]. *Beszédkutatás* 2010: 184–198.

Laczkó, M. In press. A beszéd jellemzői spontán társalgásban és tanórai megnyilatkozásokban [Characteristics of Speech in Spontaneous Conversation and Classroom Statements].

Libárdi, P. 2015. Megakadásjelenségek 17 éves diákok spontán dialógusaiban [Interruption phenomena in spontaneous dialogues of 17-year-old students]. In: S. Bátyi & M. Vigh-Szabó (eds.), *A nyelv – rendszer, használat, alkalmazás. Pszicholingvisztikai Tanulmányok V* [The language - system, usage, application. Psycholinguistic Studies V], 141–154. Budapest: Tinta Könyvkiadó.

Postma, A., H. Kolk & D.-J. Povel. 1990. On the relation among speech errors, disfluencies and self-repairs. *Language and Speech* 33(1): 19–29. https://doi.org/10.1177/002383099003300102

Silverman, S. W. & N. Bernstein Ratner. 1997. Syntactic complexity, fluency and accuracy of sentence imitation of adolescents. *Journal of Speech, Language and Hearing Research* 40(1): 95–106. https://doi.org/10.1044/jslhr.4001.95

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

Vallent K. 2009. A spontán beszéd sajátosságainak tükröződése középiskolások fogalmazásaiban [Reflection on the characteristics of spontaneous speech in the formulation of high school students]. Ph.D. dissertation, Eötvös Loránd University (ELTE), Budapest, Hungary.

Self-monitoring in children's speech

Judit Bóna
Department of Applied Linguistics and Phonetics, ELTE Eötvös Loránd University, Budapest, Hungary

During spontaneous speech, speakers monitor their own speech, and they often correct their speech errors. Self-monitoring can be covert and overt. In case of covert monitoring, the reparandum is not articulated, filled pauses and repetitions may refer to it. In case of overt monitoring, the speech error and its repair both appear in the speech.

Deciding whether a given filled pause or repetition are indeed covered repairs, is not always obvious since both carry several other functions in speech: gaining time for speech planning and word activation. In addition, filled pauses may indicate the intention to speak or to take turns.

Overt speech errors and repairs clearly express self-monitoring. There are two main repair-strategies during speaking: speakers prefer either fluency or accuracy (Seyfeddinipur et al., 2008; Nooteboom & Quené, 2017). If the speaker prefers fluency, they speak as long as they find a plan for correction. This means that they can even continue the original utterance even after having articulated the error. If the speaker prefers accuracy, they try to interrupt the reparandum as soon as possible, independent of the time they need for correction. The two strategies are characterized by different durational patterns and other acoustic-phonetic properties.

This presentation analyses the characteristics of self-monitoring of children of various ages, and and compares them to those of adults' self-monitoring. Since covert repairs are difficult to examine, therefore only self-repairs of overt errors will be analysed.

The main questions of the analysis were the following: 1) What durational patterns characterize the monitoring processes in the analysed age groups? 2) How does the dichotomy of fluency vs. accuracy occur in the self-repairs of the different age groups? Is there a difference in self-repair strategies between the examined age groups?

According to the hypotheses 1) younger children interrupt their erroneous utterances earlier than the older ones, but 2) they need longer time to correct the errors.

For the analysis, spontaneous speech samples were analysed in three age groups: in 5, 9, and 20 years of age. Speech samples were chosen from the GABI Child Language and Speech Database and Information Repository (Bóna et al., 2014) and BEA Hungarian spoken language database (Gósy, 2012). In the speech productions, self-repairs were annotated and error-to-cutoff times and editing phases were measured. The types of repairs and the repair strategies were also analysed. Finally, the data of age groups were compared.

Expected results contribute to a more accurate understanding of the age characteristics of speech production processes.

Acknowledgement

This research was supported by the Hungarian National Research, Development and Innovation Office of Hungary, project No. K-120234.

References

Bóna, J., A. Imre, A. Markó, V. Váradi & M. Gósy. 2014. GABI – Gyermeknyelvi beszédAdatBázis és Információtár [GABI - Children's Speech and Informational Database]. *Beszédkutatás* 22: 246–251.

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

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

Seyfeddinipur, M., S. Kita & P. Indefrey. 2008. How speakers interrupt themselves in managing problems in speaking: Evidence from self-repairs. *Cognition*, 108(3): 837–842. https://doi.org/10.1016/j.cognition.2008.05.004

Speech rate and pausing in school children's speech

Tímea Vakula and Éva Szennay
Faculty of Humanities, ELTE Eötvös Loránd University, Budapest, Hungary

Introduction

Different speech situations require different cognitive efforts, so speakers have different planning difficulties (Krepsz, 2016). These difficulties are visible also in temporal characteristics (Goldman-Eisler, 1968; Kowal et al., 1975). According to the literature, frequency and duration of pauses and speech situation show a correlation. Characteristics of children's pausing have been already analyzed in many international researches, considering children's age and the type of speech as a guiding factor. The percent of nursery school children's and school children's pausing was significantly higher in story-telling (about a picture) than in participating in a conversation (Deputy et al., 1982). When they had to retell a story verbatim, school children's pauses were significantly longer than when their task was only to summarise the gist of the story (Schönpflug, 2008). The duration of silent pauses decreased significantly between the age of 4 and 8 (Singh et al., 2007).

In the past years the pausing of native Hungarian nursery school and school children have been analyzed via sound recordings that consist of greater amount of spontaneous speech. During the school years changes happen in the strategy of pausing. Due to their increasing speech experiences, children start organizing the temporal patterns of their speech more economically.

According to the experiences of past researches, there are great differences in the continuance of spontaneous speech of nursery school children and school children, too. The goals of this research: (i) to compare the frequency, duration and distribution of silent pauses in nursery school children's and school children's speech (who belong to different age groups); (ii) to compare the speech rate of different aged children; and (iii) to analyze the effects that speech type has on the listed factors.

The hypotheses of this research: (i) Younger children held pauses in a greater rate and in a longer period than their older peers. (ii) Younger children's temporal rates are lower than their older peers'. (iii) Speech type has an influential effect on the analyzed factors in the three age groups.

Methodology

30 children participated in this research from GABI database (Bóna et al., 2014). There were ten 5 years old nursery school children and 20 school children, ten of them were 7 years old and the other ten were 9 years old. Children were chosen randomly from the age zones, the number of males and females were equal. In the corpus every child's native language was Hungarian. Children did not have any intellectual and hearing problems.

The children's speech in the database includes two types. First, real spontaneous speech was recorded in an interview situation, in which children talked about what kind of toy, film or book they spend their free time with. The second was a story retelling task in which the interviewer read them the same short story; that they had to listen to carefully and retell after hearing.

These sound recordings, which are the base of this study, were annotated in speech section level by using Praat software (Boersma & Weenink, 2019). Then we got the duration of pauses, the duration of speech sections and the number of syllables in an automatic way.

We calculated the proportion of pauses (measured in percent) correlated to the duration of speech, the frequency of pauses normalized to 200 syllables (Campbell & Hill, 1994), and the speech rate in syllables/second. Data were compared in the 3 age groups and in the 2 speech tasks. To the statistical analysis SPSS 23 statistic program was used, and the level of significance was 95%.

Results

The slowest speech rate was produced by a 7 year old boy, during the story retelling task, whose rate was 1.03 syllables/second. The fastest speaker was a 9 year old girl with 3.39 syllables/second in the interview task. Considering the average speed of spontaneous speech it is clear that with the increasing age, children articulate more sounds; but the individual differences are remarkable. There is no linear development in story retelling.

The time structure of the recordings was also analyzed. Considering children's speech sections and pauses as 100%, the average rate of pauses for 5-year-olds was 23.8% in the interview task and 38.5% in the store retelling task. For 7 year-old school children these average rates were 27.2% and 27.2%, respectively while for 9-year-olds the rates were 20.7% and 30.2%, respectively.

Individual differences were also significant: the lowest rate of pausing was 8.7% (produced by a 9 year old girl), and the highest one was 65.2%

(produced by a 5 year old boy). The duration of documented pauses were also defined. The shortest pauses appeared in 9 year old girls' story retelling (745 ms on average), whereas 5 year old boys produced the longest ones also in story retelling (1095 ms on average).

Our analysis consists of data referring to the rates of silent and filled pauses produced by children. Speech samples that include filled pauses in higher rates may seem more disfluent than speech samples that include only silent pauses. Filled pauses appeared in similar rates; however, individual differences were great. Almost 1/3 part of the pauses were filled pauses in more cases, nevertheless numerous children held no filled pauses at all.

Frequency of pauses was also analyzed. On average, 9 year old children held pauses the less frequently in spontaneous speech (13.38/syllables), and 5 year old children held pauses the most frequently in story retelling (4.93/syllables).

Summary

The first hypothesis is partly proved. Although the rate of pauses do not show significant changes with increasing age; older children held pauses less frequently that realize in a shorter time than in case of their younger peers. The second hypothesis is proved. With increasing age, children speak faster in interview situations, nevertheless individual differences are significant; and linear development cannot be revealed in story retelling. The third hypothesis is partly proved. Tendency-like differences can be observed between the two speech types. Hopefully the result of this study broadens the current knowledge about typically developing children's vernacular learning process.

Acknowledgement

This research was supported by the Hungarian National Research, Development and Innovation Office of Hungary, project No. K-120234.

References

Boersma, P. & D. Weenink. 2019. Doing phonetics by computer (version 6.0.46). http://www.praat.org/ (accessed 9 January 2019).

Bóna, J., A. Imre, A. Markó, V. Váradi & M. Gósy. 2014. GABI – Gyermeknyelvi beszédAdatBázis és Információtár [GABI - Children's Speech and Informational Database]. *Beszédkutatás* 22: 246–251.

Campbell, J., D. Hill & M. Driscoll. 1991. Systematic Disfluency Anlaysis: Using SDA to determine stuttering severity. Poster presented at The Annual Convention of the American Speech-Language-Hearing Association, 22–25 November 1991, Anahimeim, CA.

Deputy, P. N., H. Nakasone & O. Tosi. 1982. Analysis of pauses occurring in the speech of children with consistent misarticulations. *Journal of Communication Disorders* 15(1): 43–54. https://doi.org/10.1016/0021-9924(82)90043-0

Goldman-Eisler, F. 1968. *Psycholinguistics: Experiments in spontaneous speech*. London & New York: Academic Press.

Kowal, S., D. C. O'Connell & E. J. Sabin. 1975. Development of temporal patterning and vocal hesitations. *Journal of Psycholinguistic Research* 4(3): 195–207. https://doi.org/10.1007/BF01066926

Krepsz, V. 2016. Fonetikai hasonlóságok és különbözőségek a beszédtípusokban [Phonetic similarities and differences in speech types]. In: J. Bóna, (ed.), *Fonetikai Olvasókönyv* [Phonetic Reading Book], 175−188. Budapest: ELTE Fonetikai Tanszék.

Schönpflug, U. 2008. Pauses in elementary school children's verbatim and gist free recall of a story. *Cognitive Development* 23(3): 385–394. https://doi.org/10.1016/j.cogdev.2008.05.002

Singh, L., P. Shantisudha & N. C. Singh. 2007. Developmental patterns of speech production in children. *Applied Acoustics* 68(3): 260–269. https://doi.org/10.1016/j.apacoust.2006.01.013

Temporal aspects of disfluencies in picture-elicited story telling before and after intervention during the dynamic assessment of children's narrative skills

Ágnes Jordanidisz [1], Orsolya Mihály [2] and Judit Bóna [3]
[1]NILD Hungary, Budapest, Hungary
[2]Áldás Utcai Általános Iskola, Budapest, Hungary
[3]Department of Applied Linguistics and Phonetics, ELTE Eötvös Loránd University, Budapest, Hungary

Creating narratives requires the simultaneous activation of many linguistic and cognitive abilities. The coherence of successive thoughts cannot be created without cohesive linguistic elements (such as references to person, space, time or text itself, or ellipse). The use of textual elements requires the appropriate functioning of the working memory, but it also requires linguistic and meta-language awareness, which presupposes the existence of a certain level of schema learning and categorization skills. Narratives are generally independent of situational context and circumstances, i.e. the speaker does not use references to the physical space and time of the pronounced utterance, but must use the frame of reference in the narrative framework. This requires the ability to move the perspective, the ability of consciousness, and the ability of attentional engagement (Peterson et al., 1999; Karmiloff & Karmiloff-Smith, 2001; van Oers, 2007). In addition, the size of the active and passive vocabulary and the expressive and receptive grammatical abilities of the child also affect the level of structuring and complexity of the produced narrative. Since children's narrative skills predict later academic success (Feagans, 1982; Paul & Smith, 1993), it is important to assess it prior to the onset of instructional schooling.

Dynamic assessment offers a more holistic picture of children's real abilities than static testing, since it measures not only the current state and achievement, but also the learning potentials of the subjects as they respond to the intervention of the examiner. It is necessary to assess the initial state; calculating the gain after the "learning phase"; defining the cognitive dysfunctions and the metacognitive aspects of individual thinking, etc. (Karpov & Tzuriel, 2009).

In Hungary, one year before entering elementary school, kindergarten children are subject to compulsory native language assessment, in which their vocabulary, their grammar knowledge, their speech processing (e.g. SZÓL-E ?, Kas et al., 2012) are assessed. However, no testing procedure has been developed and implemented so far to assess the level of oral narrative production in Hungary, although kindergarteners should receive intervention before the onset of instructional learning—on the basis of the test results—to prevent later academic failures. Therefore, the development of a dynamic assessment of narrative skills has been initiated by the ELTE Child Language Research Team. The dynamic assessment uses picture-elicited story-telling to measure the narrative skills of Hungarian kindergarteners. The protocol of the assessment covers the scoring procedure and the recommended interventions during the learning phase if the child requires assistance to comprehend either the drawing or the content of the picture-series. The aim of the present research was to disclose the benefits of the interactions of the intervention provided by the assessor on the fluency factors of the story-telling. We hypothesized that 1) the fluency will develop as a result of the facilitating questions in the learning phase of the dynamic assessment and there will be fewer disfluencies in the repeated story-telling even though the complexity of it will be higher than the first picture-elicited narrative; 2) the tempo will accelerate after the learning phase.

16 Hungarian-speaking kindergarteners, 8 girls and 8 boys (6 years of age) participated in the research. Their first and second narratives based on a picture series were recorded along with the interactions initiated by the assessor in between them. The interventions aimed at the pragmatic awareness of the children. The recorded material was analyzed statistically (SPSS 20) with regards to the frequency and types of disfluencies, speech and articulation tempo, and the length of the narratives according to the number of the syllables.

We also scored the complexity of the first and the repeated story-telling based on the same picture series according to the protocol in three areas: 1) linguistic complexity on word and sentence levels, 2) cohesion, and 3) pragmatic awareness.

The results showed that significant difference between the first and second narratives was shown only in the length of the narratives, i.e. in the number of the syllables and also in the articulation tempo. It

means that children used the same types of disfluences with similar frequency in the first as in the second story-telling. However, it does indicate that children benefitted from the intervention, since the complexity and the length of the narrative increased as a result of it.

As a conclusion, the present research disclosed the benefits of using dynamic method in the assessment of kindergarteners' narrative skills. The gain was realized statistically in producing longer narratives as well as in the articulation tempo, which indicates that planning needed less effort after becoming more acquainted with the story. After further research, it might become a useful tool of assessment which provides practical advice for intervention to develop children's narrative skills.

Acknowledgement

This research was supported by the Hungarian National Research, Development and Innovation Office of Hungary, project No. K-120234.

References

Feagans, L. 1982. The development and importance of narratives for school adaptation. In: L. Feagans & D. Farran (eds.), *The language of children reared in poverty*, 95–116. New York: Academic Press.

Karmiloff, K. & A. Karmiloff-Smith. 2001. *The developing child. Pathways to language: From fetus to adolescent*. Cambridge, MA, US: Harvard University Press.

Karpov, Y. V. & D. Tzuriel. 2009. Dynamic assessment: Progress, problems, and prospects. *Journal of Cognitive Education and Psychology 8*(3): 228–237. https://doi.org/10.1891/1945-8959.8.3.228

Kas B., J. Lőrik, R. Molnárné Bogáth, A. Szabóné Vékony & N. Szatmáriné Mályi. 2012. *SZÓL-E? – Szűrőeljárás az óvodáskori logopédiai ellátáshoz* [SZÓL-E? - Screening procedure for preschool speech therapy]. Székeskesfehérvár, Hungary: LogoTech+ Kft.

Paul, R. & R. L. Smith. 1993. Narrative skills in 4-year-olds with normal, impaired, and late-developing language. *Journal of Speech, Language, and Hearing Research* 36(3): 592–598. https://doi.org/10.1044/jshr.3603.592

Peterson, C., B. Jesso & A. McCabe. 1999. Encouraging narratives in preschoolers: An intervention study. *Journal of Child language* 26(1): 49–67. https://doi.org/10.1017/S0305000998003651

van Oers, B. 2007. Helping young children to become literate: the relevance of narrative competence for developmental education. *European Early Childhood Education Research Journal* 15(3): 299–312. https://doi.org/10.1080/13502930701679718

Author index

<This page intentionally left blank.>

ISBN: 978-963-489-063-8