The 3rd International Symposium on Linguistic Patterns in Spontaneous Speech (LPSS2019)
Institute of Linguistics, Academia Sinica, Taipei
November 21-22, 2019

A Comparison of Filled Pauses in Scripted and Non-scripted Spontaneous Speech

Ralph L. Rose

Waseda University
rose@waseda.jp

Abstract

Television and film productions are heavily scripted, but intend to portray speech as unscripted within the fiction of the dramatic universe they depict. Previous evidence (Quaglio, 2009) suggests however, that various lexical features of speech occur in such scripted spontaneous speech differently than they do in actual spontaneous speech. The present study is a comparison of the occurrence of filled pause disfluencies (in English, *uh* and *um*) in scripted spontaneous speech and actual spontaneous speech, to see if the basic usage patterns are similar. Using the English-Corpora.org web site interface, filled pauses were examined in three corpora (spontaneous speech, TV transcripts, and movie transcripts) in terms of their basic frequency of occurrence, their *um*:*uh* ratios, and their structural distribution with respect to sentence boundaries. Each was also evaluated in terms of how they shifted over time. Results show that the disfluency patterns of scripted spontaneous speech are similar in many ways to that of actual spontaneous speech. The frequency of filled pauses is similar to that shown in other major corpora and the *um*:*uh* ratio also replicates a trend observed in other work (Wieling et al, 2016; Fruehwald, 2016) suggesting an ongoing shift toward the use of *um* over *uh* but with television and film speech patterns lagging that of society.

1. Introduction

Dramatic productions in television and movies aim to depict characters communicating with each other as if their speech were spontaneous. Although actors are reciting scripted language, they aim to produce speech that does not show some of the typical characteristics of read speech (cf. [1-2]). Professional actors can therefore give the appearance that they are speaking spontaneously. One identifying feature of spontaneous speech is the presence of disfluencies such as long silent pauses, filled pauses (e.g. *uh*/*um* in English), prolongations (e.g. *and so—*), and repairs (e.g. *cut the blue I mean red wire*). Thus, actors may make use of such disfluencies to make their scripted speech seem unscripted.

On the other hand, television and film production takes place under quite different conditions than would the spontaneous speech conditions they seek to depict (cf. [3]). In particular, producers may be under pressure to have actors use certain prosody or phraseology that is critical to the larger story, or they may have to fit the final production within a specified time frame (e.g., a thirty-minute prime-time slot, less time for sponsors' advertisements). Hence, the use of disfluencies might be influenced in non-authentic ways.

Thus, one question that may be considered is whether the use of disfluencies in television and film production actually pattern after the occurrence of such disfluencies in actual spontaneous speech. And furthermore, whether known patterns of change in disfluency usage over time are also reflected in television and film speech.

The present paper considers these two questions while focusing specifically on the use of filled pauses in an actual spontaneous speech corpus compared to the use of filled pauses in television and film productions over the past few decades. The remainder of the paper is organized as follows. The next section reviews the background literature on filled pauses in spontaneous speech as well as some argumentation for the use of simulated spontaneous speech as a useful speech object. After that, the corpora and investigative method applied to these corpora are described followed by the obtained results. Finally, these results are discussed together with implications for further research.

2. Background

2.1. Television and Film transcripts as linguistic objects

Dramatic productions in television and film are typically highly scripted events, with not just words and sentences decided in advance, but often even prosody as well as gestures and movements. In this respect, it is questionable whether they can be regarded as spontaneous speech. In at least one sense, they can be regarded as samples of spontaneous speech because in the fictional world of the production, speech occurs as if it is spontaneous, and other characters in the story respond to it as if it was produced spontaneously (cf. [3]).

In this sense, then, television and film transcripts can arguably be used as a source for examination of linguistic patterns and may be used by literature analysts as well as language teachers (cf. [4-5]) to the extent that they actually reflect the various linguistic patterns of spontaneous speech—or at least the type of spontaneous speech that is under investigation. Indeed, the speech events which occur in television and film production are known to undergo a certain amount of normalization ([3]). Quaglio [6], for instance, observes that hedges and vague reference occur less frequently in transcripts of the TV series *Friends* than in a corpus of conversational speech. When it comes to disfluencies, actors may even be under pressure to avoid them entirely: For instance, the actor and comedian Bob Newhart—well-known for use of disfluent speech—said that he was asked by a producer to "run some of the lines together" to save time [7].

Hence, it is an open question to what degree disfluency patterns would be reflected in the scripted scenes of spontaneous speech found in television and film productions. A further open question would be to what degree these productions—broadcast daily and viewed by millions of viewers, sometimes repeatedly in syndication—are leading or following indicators of actual speech patterns. Hopkins, Kim,

and Kim [8] in an analysis of the influence of news programs on public perceptions of the economy suggest that there is no significant change in perceptions afterward. At best, news coverage seems to reflect public attitudes in a post-hoc manner. In the same, way, perhaps it is likely that the speech found in dramatic television and film productions follows rather than leads public speech patterns. This is the basis for one of the research questions of the present study: Do disfluency patterns found in the speech of scripted television and film productions follow or lead (or neither) that of spontaneous speech?

2.2. Filled pauses in spontaneous speech

In English, the most common filled pauses take one of two closely related phonemic forms: a centralized vowel [əː] or the same with nasal closure [əːm] [9-13] although there is variation in the vowel across dialects [14]. Orthographically, these two forms are typically represented as *uh* and *um* respectively in North American English print (in post-vocalic *r*-less regions such as the UK, they are more likely rendered as *er* and *erm*, respectively).

Most researchers have treated these two variants as identical beyond their differing phonemic shapes, but some have sought to distinguish between the two. Several researchers [12, 15-18] have observed that silent pauses following *um* are longer than those following *uh*. Fox Tree [19] observed in both English and Dutch that listeners recognized words in speech following an *uh* better than those following an *um*. Based on this, Clark and Fox Tree [12] hypothesize that the two variants are used by speakers differentially, as follows: *uh* is used by speakers to signal that they anticipate a short(er) delay in their speech production while *um* is used to signal their anticipation of a long(er) delay. (Note that this hypothesis is not without controversy: Corley & Stewart [20] and Finlayson [21] argue against this intentional and designed 'use' of filled pauses in spontaneous speech.)

2.2.1. Rate of occurrence

The reported rate of occurrence of filled pauses varies highly among studies. Figure 1 shows the frequency of occurrence in several corpora of English speech (data from Tottie [22]).

Differences between corpora have been hypothesized to be due to gender (cf. [13]) and age (cf. [23]) differences in filled pause use and differences in these groups' representation in each corpus. Nonetheless, in nearly every speech corpus, filled pauses appear as one of the most frequent tokens.

2.2.2. Um:Uh ratio variation

As Figure 1 also illustrates, there is wide variation in the occurrence of the two filled pause variants relative to each other. At one end, the Switchboard Corpus shows an um:uh ratio of approximately 1:5. But on the other hand, the Louvain Corpus of Native English Conversation shows a ratio of nearly 2:1. As noted above, this variation may be driven, in part, by gender and age differences in usage and corpus representation. Wieling et al [24] and Fruewald [25] have investigated the um:uh ratio in English and other Germanic languages and observed a common trend among all the languages where the use of *um* is increasing relative to *uh*, and this change is being led by younger people and females. Figure 2 shows the English data, taken from the Philadelphia Neighborhood Corpus [25].

Although the data in their study clearly show an ongoing change in the um:uh ratio modulated by age and gender, it is important to note that only the youngest have flipped the ratio on average to use more *ums* than *uhs* (females born after 1975 and males born after 1995).

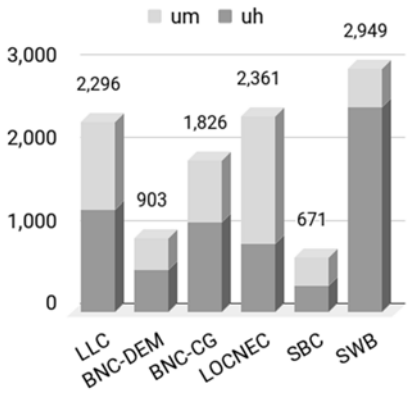

Figure 1. *Frequency of filled pauses per 100K words in various corpora: London-Lund Corpus (LLC), Demographic spoken portion of British National Corpus (BNC-DEM), Context-governed spoken portion of British National Corpus (BNC-CG), Louvain Corpus of Native English Conversation (LOCNEC), Santa Barbara Corpus (SBC), and Switchboard Corpus (SWB). Data from Tottie [22].*

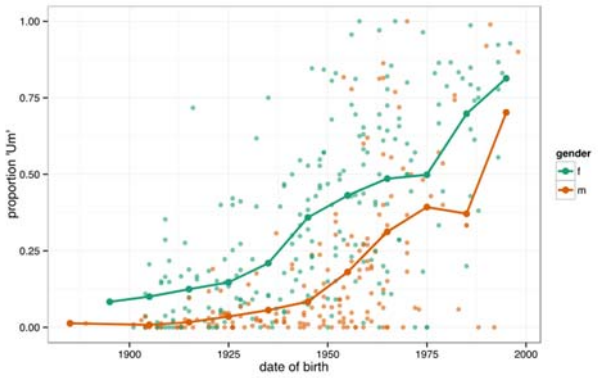

Figure 2. *Proportion of um use in Philadelphia Neighborhood Corpus (figure reproduced from Fruehwald [25] with permission).*

2.2.3. Structural distribution

Another observation on the distribution of filled pauses is that they occur somewhat differentially with respect to discourse structure. Several studies show a trend wherein filled pauses occur more frequently at major rather than minor discourse boundaries [26-27]. Furthermore, in comparable work, Rose [28] observed that filled pauses at clause boundaries were more likely to be *um* than *uh*, though this trend was stronger for monologic speech than for conversation.

3. Corpus investigation

Based on the above evidence, the present study seeks to compare the occurrence of filled pauses in spontaneous speech to that in scripted spontaneous speech as represented by speech in television and film productions. In particular, the following empirical research questions are explored:

1. What is the rate of occurrence of filled pauses and how does it change over time?
2. What is the *um*:*uh* ratio of filled pauses and how does it change over time?
3. What is the distribution of *um* and *um* at structural positions and how does it change over time?

In addition to these questions is the question, identified above, as to whether scripted television and film speech might lead or follow changes in actual spontaneous speech.

The following subsections describe the corpora and the method used to evaluate these questions and then the results of the investigation.

3.1. Corpora

The corpora used in the present study are among those available at the English-Corpora.org web site at Brigham Young University and are described as follows.

- Corpus of Contemporary American English (COCA [29]): The spoken portion of this corpus is used, consisting of transcripts of spontaneous speech from US TV news and talk shows, from 1990 to 2017. (117 million words with 11,978 ums/uhs)
- The TV Corpus (TV [30]): The US-Canadian portion of the corpus is used, consisting of transcripts of dramatic television shows originally broadcast between 1950 and 2018. (326 million words with 580,952 ums/uhs)
- The Movie Corpus (Movie [31]): The US-Canadian portion of the corpus is used, consisting of transcripts of films released between 1930 and 2018. (199 million words with 211,997 ums/uhs)

COCA is used in the present study as a reference corpus for spontaneous speech trends in disfluencies. The TV and Movie corpora are then used comparatively to evaluate how scripted spontaneous speech differs. The three corpora do not consistently distinguish speaker boundaries, nor give any clear speaker meta-information, so such things as age and gender effects or any hypotheses related to individual speakers cannot be evaluated directly. However, it may be safe to assume that comparing an earlier portion of each corpus to a later portion of each approximates a comparison of one generation to a later generation, respectively. This is based on an assumed analogy: that the speech of today's older people is to that of today's younger people as the speech exhibited in older transcripts (e.g. pre-1970) is to that exhibited in younger transcripts (e.g. post-1970).

3.2. Method

The corpus investigation took place using the tools available through the English-Corpora.org web interface using an account with a free license. This includes the capability to search on individual words (i.e., *uh* and *um*), breakdown such searches with respect to year, and also download a minimal amount of keyword-in-context (KWIC) data for closer analysis. A paid license account is required to download each corpus in its entirety for sophisticated off-line analysis, But with a free account, KWIC data of up to 1,000 hits at a time can be viewed. This provided a sufficient number of cases for the basic analysis reported in the present research. In order to examine the structural distribution of filled pauses, it is assumed that filled pauses in the transcripts which are capitalized (e.g., *Uh*, *Um*) occur only sentence-initially and therefore can be used to distinguish between sentence boundary (capitalized) and sentence non-boundary (non-capitalized) instances of filled pause use.

In all cases, the searches were limited to *uh* or *um* for general frequency searches or capitalized and non-capitalized variants of these for the boundary–non-boundary searches described above. Other orthographic forms of filled pauses can be found in the corpora (e.g. *uhh*, *umm*, *er*, *erm*, *ah*), but these are quite rare compared to *uh* and *um* (less than 1%) and some of them actually denote non-filled pause phenomena (e.g. *ah* as a sign of surprise or realization or *ER* as a colloquial abbreviation for "emergency room").

Not all instances of *uh* and *um* are necessarily filled pause uses. Examination of KWIC results suggests that some are actually part of other expressions (e.g. *uh huh* or *um hmm*), repetitive filled pause sequences (*uh uh uh*…) or grunt-like interjections (*um!*). In the TV and Movie corpora, the KWIC results show that these cases are very low (nearly 0% in TV and 2.5% in Movie), but are significant in the COCA corpus (14.7%). Yet, where they do occur, the occurrence does not strongly favor either *uh* or *um*, nor does it favor any particular time period. This means this trend would have an effect only on the basic frequency count results, with the COCA data showing a slight over-count. The *um*:*uh* ratios should not be affected and the structural distribution analysis will directly evaluate the KWIC data, so the deviant cases will be removed.

In any case, the underlying analytical question being looked at here is how these frequencies and ratios are changing over time. Thus, as long as there is no *um*:*uh* or time-related bias in the deviant cases, the results should not be influenced.

The statistical analyses reported below were performed in `R` (version 3.3.2) using linear regression (`lm`) and Mantel-Haenszel χ^2 test (`mantelhaen.test`) with an alpha level of 0.05. Tests were performed mainly using the data from 1990 to 2017—the range of years for which there is data from all three corpora. For the linear regression tests, time was treated as scalar variable; for the χ^2 tests, time was treated as a two-level categorical variable (pre-2000, post-2000).

3.3. Results

The frequency of filled pauses is shown in Figure 3. The results show that filled pauses are used in the TV and Movie corpora at a rate significantly higher than in COCA [$t(78)=3.3$, $p=0.001$]. However, the TV and Movie rates are in the same broad range as the corpora illustrated in Figure 1. Furthermore, this rate is rapidly increasing in recent years, despite the fact that the use of filled pauses in COCA is relatively stable [$t(78)=3.4$, $p<0.001$]. The TV corpus shows a very high rate of use from 1950 to about 1975. However, this is likely due to limited data: There are not so many scripts from these years and as the line shows, there is large variation from year to year during this period.

Figure 4 shows the *um*:*uh* ratio for filled pauses. Results show that all three corpora exhibit the same trend in recent years toward greater proportional use of *um* to *uh* [$t(78)=14.1$, $p<0.001$]. While this is consistent across all three corpora, it is most pronounced in the COCA corpus [$t(78)=5.9$, $p<0.001$], with the other two lagging somewhat behind. The COCA data is further consistent with the Wieling et al [24] and Fruehwald [25] observations showing that a preference for *um* over *uh* is appearing in recent years.

Figure 5 shows the distribution of filled pauses relative to sentence boundaries. The results show that COCA shows a clear, consistent pattern across time [$\chi^2(1) = 11.4, p < 0.001$]: The open vowel filled pause, *uh*, is more frequently used in sentence non-boundary positions, while the nasal filled pause, *um*, is more commonly used at sentence boundary positions. This is consistent both before the year 2000 as well as after. However, both the TV and Movie corpora show different patterns from COCA as well as from each other. The TV corpus

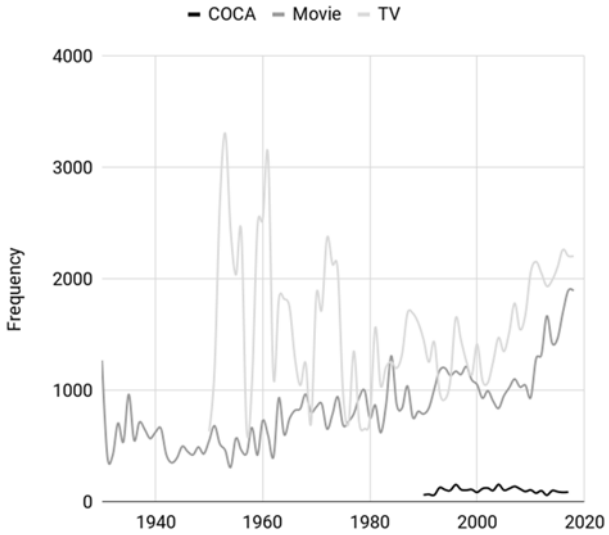

Figure 3. *Frequency (per million words) of filled pauses over time, from 1930 (TV), 1950 (Movie) and 1990 (COCA) to 2018.*

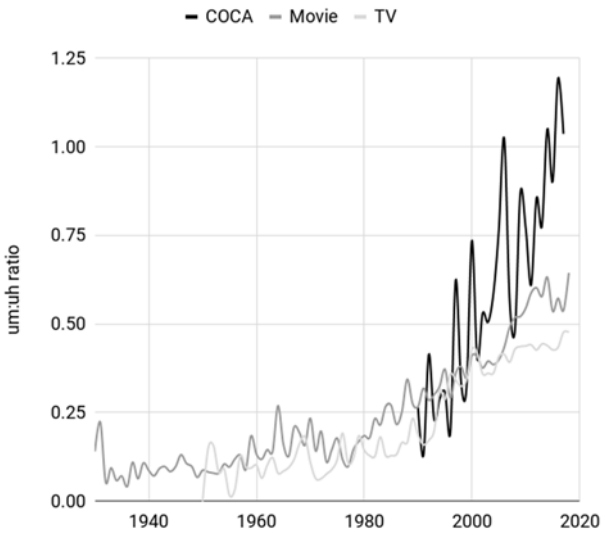

Figure 4. *Um:Uh ratio for filled pauses over time, from 1930 (TV), 1950 (Movie) and 1990 (COCA) to 2018.*

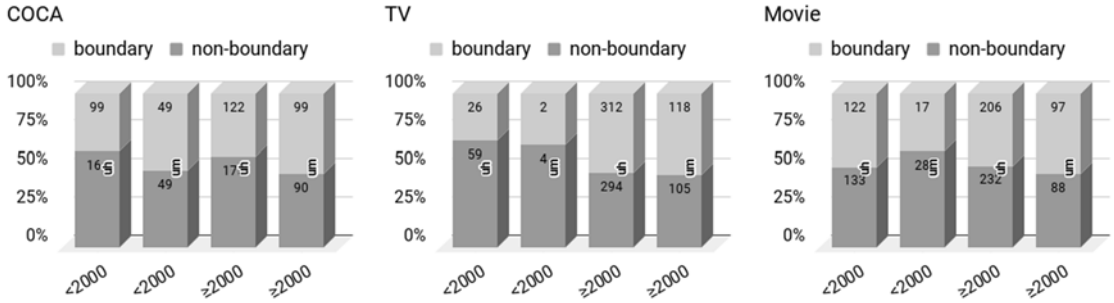

Figure 5. *Proportion of filled pauses (uh, um) at sentence boundary and non-boundary positions in early (<2000) and late (≥2000) years.*

suggests that *um* and *uh* are similar, but that their use changed over time [$\chi^2(1) = 16.5$, $p < 0.001$]: Before 2000, both were preferentially used at non-boundary positions but after 2000, at boundary positions. In the Movie corpus on the other hand, *uh* has remained consistent over time, with a roughly even distribution between boundary and non-boundary positions, but *um* has shifted from a preference for non-boundary before 2000 to a preference for boundary positions after 2000 [$\chi^2(1) = 22.4$, $p < 0.001$].

4. Discussion

The present study has sought to evaluate whether scripted spontaneous speech shows similar filled pause usage patterns to authentic spontaneous speech by comparing corpora of transcripts of dramatic television and film productions to a corpus of spontaneous speech in English. Three specific previously observed trends in filled pause use were evaluated.

Results show that the TV and Movie corpora show filled pause usage rates that are significantly higher than that of the COCA corpus. However, their rates are comparable to that of other large corpora (see Figure 1). Thus, it may be the case that COCA is the outlier here. The speech samples that are used in the COCA corpus are from TV news and talk shows and feature people who are professional speakers. They use filled pauses, but perhaps at a rate diminished over the speakers in the corpora in Figure 1 who are more likely to have no special training in public speech.

This may raise the question whether COCA is an appropriate corpus to use as a representative of spontaneous speech, given its make-up of a larger proportion of speech by professional speakers. However, note that the frequency of filled pauses in COCA is at just the outside range of the corpora shown in Figure 1, comparable to the pattern seen in the Santa Barbara Corpus (SBC). SBC [32] consists of completely unscripted, unprepared speech, recorded ambiently by a recorder placed in various environments. Thus, it is apparent that there is wide variation in the frequency of filled pauses in various corpora. As such, any conclusions here based on basic frequency counts must be tentative. The present work therefore emphasizes the ratio of *um* to *uh* usage as well as the changes in the frequency over time.

Regarding this last point, an interesting trend is the increase in filled pause use in the TV and Movie corpora beginning in about 2000. It is not immediately clear what might be driving this trend. One possibility is that it might be reflecting the rapid growth of personal video-sharing: As this social trend increases, depictions of them in TV and Movie may be increasing, and concurrent with it, the spontaneous speech style found in such impromptu videos. This, however, is just speculation and needs further testing for confirmation.

Another observation concerns the *um*:*uh* ratio. The COCA results replicate previous work and suggest that TV and Movie speech are reflecting the trend toward increased use of *um* relative to *uh*, and even approaching the same magnitude. While the results for the TV and Movie corpora follow this pattern, they are somewhat diminished. This suggests that TV and Movie speech are not helping to influence this change but are merely reflecting this change post hoc.

Finally, the structural distribution results are somewhat confounding. Neither the TV nor Movie corpus pattern after the COCA corpus in terms of the use of filled pauses at sentence boundary and non-boundary positions, nor are their patterns similar to each other. One possible explanation for the Movie corpus might come from the *um*:*uh* ratio studies: Wieling et al [24] and Fruehwald [25] conjecture that the reason for the increased use of *um* in recent years is perhaps sociolinguistic: They suggest that *um*—given its high co-occurrence with longer silent pauses—is gaining prominence as an overt politeness signal. If this explanation is plausible, then perhaps the Movie corpus is reflecting it. The COCA corpus (which actually is a corpus of spontaneous speech) does not reflect this change, but the explanation for this, might be the same as above: Perhaps the professional speakers featured in the transcripts do not have as much need for this form of politeness marker, instead using common lexical devices to accomplish the same purpose. This explanation, however, does not account for the unusual pattern shown in the TV corpus. Clearly, more work would be needed to clarify these observations.

5. Conclusions

The results show here that the disfluency patterns of scripted spontaneous speech are similar in many ways to that of actual spontaneous speech: The basic frequency patterns of occurrence are quite similar between the two, even showing similar trends over time. However, the discrepancy comes at the level of structural distribution. These findings may be useful for a better understanding of how scripted spontaneous speech may inform normal speech studies or language teaching as well as how (or whether) such corpus data may be used as resources in machine learning applications.

6. References

[1] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech," *Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2862–2873, 2002.

[2] C. Cucchiarini, J. van Doremalen, and H. Strik, "Fluency in non-native read and spontaneous speech," *Proceedings of DiSS-LPSS Joint Workshop 2010 – 5th Workshop on Disfluency in Spontaneous Speech and 2nd International Symposium on Linguistic Patterns in Spontaneous Speech*, pp. 15–18, 2010.

[3] M. Alvarez-Pereyre, "Using film as linguistic specimen: Theoretical and practical issues," In Piazza, R., Bednarek, M. & Rossi, F. (eds.), pp. 47–67, 2011.

[4] E. Csomay and M. Petrović, "'Yes, your honor!': A corpus-based study of technical vocabulary in discipline-related movies and TV shows," *System*, vol. 40, no. 2, pp. 305–315, 2012.

[5] R. Piazza, M. Bednarek, and F. Rossi, (eds.) *Telecinematic Discourse: Approaches to the language of films and television series*. Amsterdam: John Benjamins Publishing Company, 2011.

[6] P. Quaglio, *Television Dialogue: The sitcom Friends vs. natural conversation*. Amsterdam: John Benjamins Publishing Company, 2009.

[7] B. Newhart, Interview on *The Today Show* (NBC), November 14. 2010.

[8] D. J. Hopkins, E. Kim, and S. Kim, "Does newspaper coverage influence or reflect public perceptions of the economy?" *Research and Politics*, vol. 4, no. 4, pp. 1–7, 2017.

[9] H. Maclay and C. Osgood, "Hesitation Phenomena in Spontaneous English Speech," *Word*, vol. 15, no. 1, pp. 19–44, 1959.

[10] F. Goldman-Eisler, "A Comparative Study of Two Hesitation Phenomena," *Language and Speech*, vol. 4, no. 1, pp. 18–26, 1961.

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

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

[13] E. Shriberg, *Preliminaries to a theory of speech disfluencies*, Ph.D Thesis. Berkeley, CA: University of California, Berkeley, 1994.

[14] R. J. Lickley, "Fluency and Disfluency" in Redford, Melissa A. (ed.), The Handbook of Speech Production. Chichester, UK: Wiley Blackwell, 445–474. 2015.

[15] H. Clark, "Managing problems in speaking," *Speech Communication*, vol. 15, no. 3/4, pp. 243–250, 1994.

[16] H. Clark and J. Fox Tree, "On thee-yuh fillers uh and um." *Language Log*, 2014. https://languagelog.ldc.upenn.edu/nll/?p=15718 (accessed June 6, 2019).

[17] T. Kendall, *Speech rate, pause and sociolinguistic variation*. Basingstoke: Palgrave Macmillan, 2013.

[18] R. Rose, *The communicative value of filled pauses in spontaneous speech*, Master's Dissertation, Birmingham, UK: University of Birmingham, 1998.

[19] J. E. Fox Tree, "Listeners' uses of 'um' and 'uh' in speech comprehension," *Memory and Cognition*, vol. 29, no. 2, pp. 320–326, 2001.

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

[21] I. R. Finlayson, *Testing the roles of disfluency and rate of speech in the coordination of conversation*, Ph.D. Thesis. Edinburgh, Scotland, UK: Queen Margaret University, 2014.

[22] G. Tottie, "Uh and Um as sociolinguistic markers in British English," *International Journal of Corpus Linguistics*, vol. 16, no. 2, pp. 173–197, 2014.

[23] H. Bortfeld, S. D. Leon, J. E. Bloom, M. F. Schober,.and S. E. Brennan, "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender," *Language and Speech*, vol. 44, no. 2, pp. 123–147, 2001.

[24] M. Wieling, J. Grieve, G. Bouma, J. Fruehwald, J. Coleman, and M. Liberman, "Variation and Change in the Use of Hesitation Markers in Germanic Languages," *Language Dynamics and Change*, vol. 6, no. 2, pp. 199–234, 2016.

[25] J. Fruehwald, "Filled Pause Choice as a Sociolinguistic Variable," *Penn Working Papers in Linguistics*, vol. 22, no. 2, pp. 42–49, 2016.

[26] M. Swerts, "Filled pauses as markers of discourse structure," *Journal of Pragmatics*, vol. 30, no. 4, pp. 485–496, 1998.

[27] M. Watanabe, "Comparison of factors related to clause-initial filler probabilities in English and Japanese," *Proceedings of International Congress for Phonetic Sciences*, pp. 2440–2444, 2019.

[28] R. Rose, (2015). "Um and uh as differential delay markers: the role of contextual factors," paper presented at *The 7th Workshop on Disfluency in Spontaneous Speech (DiSS 2015)*, 2015.

[29] M. Davies, (2008–) "The Corpus of Contemporary American English (COCA): 560 million words, 1990–present," 2008–. https://www.english-corpora.org/coca/ (accessed June 6, 2019).

[30] M. Davies, "The TV Corpus: 325 million words, 1950–2018," 2019a–. https://www.english-corpora.org/tv/ (accessed June 6, 2019).

[31] M. Davies, "The Movie Corpus: 200 million words, 1930–2018," 2019b–. https://www.english-corpora.org/movies/ (accessed June 6, 2019).

[32] J. W. Du Bois, W. L. Chafe, C. Meyer, and S. A. Thompson, *Santa Barbara corpus of spoken American English, Part 1*. Philadelphia: Linguistic Data Consortium, 2000.