University of Pennsylvania Working Papers in Linguistics

Volume 17
Issue 2 *Selected Papers from NWAV 39*

Article 2

2011

On Gender Differences in the Distribution of um and uh

Eric K. Acton
Stanford University

On Gender Differences in the Distribution of um and uh

Abstract

While the so-called “fillers” *um* and *uh* share a great deal in the way of interpretation, association, and usage, they are far from perfect substitutes. Previous corpus research, focusing primarily on British English, has identified a number of social and discursive factors with which filler usage can vary, including pause length and position in an utterance and speaker age, gender, and social class (Rayson et al. 1997, Clark and Fox Tree 2002, Tottie 2011, *inter alia*). Building on such research, the present paper investigates social variation in the use of *um* and *uh* in the United States. In particular, the paper documents the results of two corpus-based investigations of women’s and men’s usage of *um* and *uh* demonstrating that, among the speakers represented in the corpora, women on the aggregate had a far higher ratio of *um* tokens to *uh* tokens (*um*/*uh* ratio) than did men. The first of the two corpora examined is a collection of 992 transcripts from three speed-dating events held for graduate students at an American university in 2005. In this corpus, women’s average *um*/*uh* ratio is more than 3.5 times that of men. An analysis of gendered filler usage in the Switchboard Corpus (SWBC) yields a similar result: women’s average *um*/*uh* ratio in the SWBC is more than 2.5 times that of men. Data from the SWBC likewise suggest that this general trend persists across age groups and major U.S. dialect regions and, furthermore, tends to hold for speakers regardless of the gender of their interlocutors. The SWBC also provides evidence suggesting that *um* is gaining currency relative to *uh*; i.e., that there is a linguistic change in progress whereby the use of *um* relative to *uh* is on the rise. It is noted that not all men and women in the corpora exhibit filler usage in line with the aggregate-level trends, and that gendered linguistic differentiation should not be assumed to be a direct reflection of gender *per se* (Eckert 1989). A thorough understanding of the dynamics of gender and filler usage calls for an examination of the meanings and associations of *um* and *uh* and of speakers’ stances, objectives, and relation to their social world.

On Gender Differences in the Distribution of *um* and *uh*

Eric K. Acton*

1 *Um* and *uh*: An Introduction

Consider the following:

(1) **Um**, I'm, I'm an artist. I'm a dancer, poet, **um**, writer, comedian. Not that, like, I say those in big letters. I just do it for fun. I think that's, that's, that's the true, the true pleasure of art is not, not to get all obsessed with, you know, **um**, just, I don't know…

(2) **Uh**, I'm, I'm an artist. I'm a dancer, poet, **uh**, writer, comedian. Not that, like, I say those in big letters. I just do it for fun. I think that's, that's, that's the true, the true pleasure of art is not, not to get all obsessed with, you know, **uh**, just, I don't know…

The two specimens above are identical except in two key respects: first, every instance of *um* in (1) has been replaced by *uh* in (2); and second, only one of the two is an excerpt from a verbatim transcription of an actual conversation between two people on a speed date. Despite the formal differences between them, I suspect that most people, at least at first glance, would regard the two passages as essentially equivalent in terms of what they communicate.

Nonetheless, the view taken in this paper is that while *um* and *uh* share a great deal in the way of interpretation, association, and usage, they are far from perfect substitutes. English speakers may, for instance, have intuitions regarding whether (1) or (2) sounds more natural, or may even imagine distinct speakers for the two examples. Moreover, previous research has revealed significant differences in the distribution of *um* and *uh*. Based on a corpus study of conversations between British adults recorded from 1961 to 1976, Clark and Fox Tree (2002) reported that although *um* and *uh* both served the discourse function of signaling the initiation of a delay in speech, pauses initiated by *um* were generally longer than those initiated by *uh*; and that *um* was more likely than *uh* to be found at intonation-unit boundaries and less likely than *uh* to be found within intonation units. Nor do the documented distinctions between *um* and *uh* pertain only to their discourse functions. In a study of the British National Corpus (BNC), Rayson et al. (1997) found that the two expressions also differ from each other in their distribution along social category lines. Specifically, the authors found in the spoken portion of the BNC that men and speakers over 34 years of age tended to say *uh*[1] more than women and speakers 34 or younger, respectively. The authors likewise reported that, when the speakers were divided into two socio-economic classes based on occupation, *um* was more prevalent among the speech of those from the higher of the two classes. In related work on the spoken BNC, Tottie (2011) reported that, on average, the use of *um* relative to *uh* in the corpus was significantly higher for women than for men, and varied inversely with age and directly with socio-economic class.

Building on such research, the present paper investigates social variation in the use of *um* and *uh* in the United States. In particular, I will present the results of two corpus-based investigations of women's and men's usage of *um* and *uh* demonstrating that, among the speakers represented in the corpora, women on the aggregate had a far higher ratio of *um* tokens to *uh* tokens than did men. The data from the second of the two corpora under investigation suggest that this general trend persists across age groups and major U.S. dialect regions and, furthermore, tends to hold for speakers regardless of the gender of their interlocutors. I will also provide evidence suggesting that *um* is gaining currency relative to *uh*; i.e., that there is a linguistic change in progress whereby the use of *um* relative to *uh* is on the rise. The paper concludes with a brief discussion of the questions raised by this research, with an emphasis on questions concerning the meaning of the two

*Special thanks to Daniel Jurafsky, Penelope Eckert, Christopher Potts, Meghan Sumner, John Rickford, and Beth Levin for their guidance and insight. For Kevin and J.C.

[1] Results in Rayson et al. were in fact reported for expressions transcribed as *er* and *erm*, which are British spellings of *uh* and *um*, respectively (Clark and Fox Tree 2002, Tottie 2011). (This, of course, is not to say that *uh*/*er* and *um*/*erm* are pronounced identically by speakers across, or even within, all dialects of English.)

fillers.

Before I proceed, a couple of notes are in order. First, I address matters of terminology. Following Clark and Fox Tree (2002), I will refer to *um* and *uh* collectively as "fillers" throughout the discussion. This is motivated entirely by expository convenience, and it should not be considered indicative of any particular view regarding the meaning or function of *um* or *uh* or whether other words should likewise be classified as fillers. In addition, I will need a consistent metric by which to assess the relative usage of *um* and *uh* for a particular speaker or set of speakers. To that end, for a given speaker or set of speakers *S*, I define the "*um*/*uh* ratio" of *S* to be *S*'s total number of tokens of *um* divided by *S*'s total number of tokens of *uh*.

Lastly, a word on gender. I wish to state from the outset that the purpose of this analysis is not to make essentialist claims about differences in men and women's speech, or the speech of members of any other social category, for that matter. Rather, generally speaking, I am concerned with observed linguistic variation along gender lines only insofar as such variation (i) may teach us something about the social landscape in which speakers participate; or (ii) suggests that the variants under consideration differ not only in form but also in meaning. Furthermore, many scholars have rightly pointed out that gender is indeed a highly nuanced social construct and should be approached as such in studies of linguistic variation (Eckert and McConnell-Ginet 1992, Fuller Medina and Roy 2010, Macaulay 1978, *inter alia*). At the same time, it is clear that the gross categories of "male" and "female" are hugely instrumental in the organization and understanding of our social world, thereby offering much to the study of meaning and meaning-making, particularly of the social kind.

2 Gender Differences in the Distribution of *um* and *uh*: Evidence from Corpus Research

The data discussed in this section come from two rather distinct corpora: the Speed Dating Corpus and the Switchboard Corpus. The two corpora, and the analyses of each, will be addressed in turn.

2.1 The Speed Dating Corpus

The Speed Dating Corpus (SDC) is a collection of audio recordings from three speed-dating sessions held for graduate students at a private American university in 2005 (Jurafsky et al. 2009). Participants wore audio recording devices during the sessions and were told that their conversations would be recorded for research. Over the course of the three sessions, audio recordings of approximately 1,100 four-minute dates were made. 992 of these recordings were later transcribed by professional transcribers. It is this set of 992 written transcripts, containing over 750,000 words, which served as the basis for the research described here.

The dating sessions were conducted in a round-robin fashion, wherein each dater had a date with each dater of the other gender. There were no same-gender dyads among the dates. The overt heterosexuality of these events no doubt made for an atmosphere in which gender and gender roles were especially pronounced.

A lexical frequency analysis of the corpus revealed that the two genders shared many of the same most commonly spoken words: on the aggregate, both men and women used *I* more than any other word, *you* second most, and *like* fifth most, with *to* and *yeah* alternating between being the third and fourth most frequent words. One pair of words, however, exhibits a sharp distributional contrast between women and men: *um* and *uh*.

Aggregating by gender we find a small but significant difference in the rate of *um* and *uh* taken together, with fillers comprising 1.06% of all of women's words, compared with 1.14% of men's ($\chi^2 = 10.70$, 1 d.f., $p = 0.001$). A far more dramatic difference, however, can be found between the two genders' *um*/*uh* ratios. According to the written transcripts, *um* was the 24th most spoken word among women, and the 43rd among men. For *uh* we have a near mirror image: it ranked 25th for men and 62nd for women. The distribution of *um* and *uh* aggregated by gender is provided below in Table 1.

As shown in Table 1, women's *um*/*uh* ratio was over 3.5 times that of men. Nor are these group-level differences attributable to a small number of daters. Table 2 reports the percentage of women and men who used *um* more than *uh*. A majority of women (79.6%) used *um* more than *uh*, while well under half of men (32.1%) did.

	Tokens of *um*	**Tokens of *uh***	**Combined**	***um*/*uh* ratio**
Women	2,814	1,263	4,077	2.23
Men	1,692	2,789	4,481	0.61

Table 1: Speed Dating Corpus: *Um* v. *uh* by speaker gender.[2]

	Speakers with more *um*s	**Speakers with more *uh*s**	**N**	**% with more *uh*s**
Women	43	11	54	79.6%
Men	18	38	56	32.1%

Table 2: Speed Dating Corpus: Speaker usage preference for fillers by gender.

As one additional check on the durability of these trends, the values from Table 1 were recalculated, this time excluding non-native English speakers, who made up 24.5% of the participants. The concern was that the disproportionately many non-native speakers among the men might explain much of the gender difference. As shown in Table 3, however, removing the participants coded in the corpus as non-native English speakers does little to narrow the gap:[3]

	Tokens of *um*	**Tokens of *uh***	**Combined**	***um*/*uh* ratio**
Women	2,674	1,171	3,845	2.28
Men	1,122	1,577	2,699	0.71

Table 3: Speed Dating Corpus: *Um* v. *uh* by speaker gender, native English speakers only.[4]

It is true that with the non-native English speakers removed, both genders' *um*/*uh* ratios increase. In the case of men, the increase is an appreciable 17.2%, from 0.61 to 0.71. That said, the difference in use between the two genders remains immense, with women's *um*/*uh* ratio still more than three times that of men.

Given the rarity of linguistic gender differences of this magnitude, I also conducted an analysis of whether *um* and *uh* were reliably transcribed in the SDC. Fifty fillers originally transcribed as *um* and 50 transcribed as *uh* were randomly selected from the corpus, and checked for fidelity vis-à-vis the audio recordings. Each of the 100 tokens to be checked was identified according to the date it occurred in and at what point in the date it was said, allowing me to blind myself to whether it was originally transcribed as *um* or *uh*. My own transcriptions of the fillers matched 91% of the time with the corpus transcriptions. Moreover, if the discrepancies in fact reflect a more general bias in the corpus transcriptions, the results of this analysis suggest that women's average *um*/*uh* ratio is even higher relative to that of men in the SDC than reported above. Among the nine discrepancies between the original transcriptions and mine, five were spoken by women.

[2] There were 68 tokens transcribed as "uhm" in the transcripts (41 for women and 27 for men), and one token transcribed as "umm" for each of the two genders. All such tokens are included in the "Tokens of *um*" column of Table 1.

[3] Classification as native English speaking in the SDC does not require a participant to be a native speaker of American English in particular. The results in Table 3 may therefore include some speakers of other varieties of English.

[4] There was a single token of *um* that could not be attributed to a particular speaker and was therefore omitted from this table.

Of those five, only one was a token that I transcribed as *uh* but had been transcribed in the corpus as *um*; the other four went in the opposite direction. Regarding the four discrepancies involving male speakers, two were transcribed in the corpus as *um*, one of which I transcribed as *uh* and the other as *so*; and the other two were transcribed in the corpus as *uh*, both of which I transcribed as *um*.

	Corpus Transcriptions			**Transcription Replication**		
	Tokens of *um*	**Tokens of *uh***	***um*/*uh* ratio**	**Tokens of *um***	**Tokens of *uh***	***um*/*uh* ratio**
Women	31	17	1.82	34	14	2.43
Men	19	33	0.58	19	32	0.59

Table 4: Speed Dating Corpus: Comparing the transcription of 100 randomly selected fillers.

Table 4 shows that the *um*/*uh* ratio among the 100 randomly selected fillers was higher for both genders for my transcriptions than for the original corpus transcriptions. The difference across the two transcriptions, however, is far greater for women (33.2%) than for men (3.1%), so that whereas women's *um*/*uh* ratio for the 100 randomly selected tokens is 3.2 times that of men based on the original corpus transcriptions, it is 4.1 times that of men based on my transcriptions. Whether or not these differences are in fact representative of a corpus-wide transcription bias, the high degree of consistency between the original transcriptions and my own suggests that the gender differences discussed above are indeed robust and great in magnitude.

Of course, if we wish to fully understand the relationship between gender and fillers, we must not be satisfied to stop at the level of abstraction we have adopted thus far. Recall from Table 2 that a nontrivial subset of both men and women went against the macro trends for their respective genders in their usage of *um* and *uh*: one in five women in the SDC actually favored *uh* over *um*, and nearly one in three men favored *um* over *uh*. For instance, one female with 119 tokens of fillers used *uh* 66% of the time,[5] and one male with 101 tokens used *um* 86% of the time. Was such variation gender-motivated? Did these participants exhibit additional language features that set them apart from other members of their respective genders? A thorough investigation of individual speakers' behavior falls outside of the scope of this paper, but will be essential to understanding the meaning and social dynamics of filler usage.

In part to underscore the robustness of the macro-level trends presented above, I now turn to an analysis of the Switchboard Corpus.

2.2 The Switchboard Corpus

The second study of filler usage was based on the Switchboard Corpus (SWBC): a database of over 2,400 telephone conversations (averaging six minutes in length) between people across the United States, recorded in 1990.[6] Although some of the conversational dyads in the SWBC are female-male, there are also hundreds of male-male (699) and female-female (651) conversations.

In analyzing the conversations in the SWBC, the first step was to determine whether the women in this corpus, like those in the SDC, had a higher *um*/*uh* ratio on average than did men. Again, we see a dramatic difference in the aggregate behavior of the two genders, as shown in Table 5. In this case, women's *um*/*uh* ratio is more than 2.5 times that of men—a factor not as great as was observed in the SDC data, but sizable nonetheless.

[5] The linguistic behavior and style of this particular speaker could make for a study unto itself. In addition to being the only one of the 54 females in the study to use the word *bitch* during a date, she accounted for 1/10 of all 30 tokens of *shit* in the SDC and 1/8 of the 24 tokens of *fuck*.

[6] Switchboard-1 Release 2. 2010. Retrieved May 18, 2011, from http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97S62.

	Tokens of *um*	**Tokens of *uh***	**Combined**	***um*/*uh* ratio**
Women	12,400	24,247	36,647	0.51
Men	8,734	45,402	54,136	0.19

Table 5: Switchboard Corpus: *Um* v. *uh* by speaker gender.

And this general pattern persisted at the dialect-region level:

Region	**Female**	**Male**	**Female ratio**
	***um*/*uh* ratio**	***um*/*uh* ratio**	**Male ratio**
Mixed	0.57	0.19	3.0
New England	0.77	0.20	3.9
North Midland	0.66	0.24	2.7
Northern	0.61	0.21	2.8
NYC	0.72	0.23	3.1
South Midland	0.39	0.15	2.5
Southern	0.41	0.18	2.3
Western	0.51	0.18	2.8

Table 6: Switchboard Corpus: *Um*/*uh* ratios by speaker gender and geographic region.[7]

As displayed in Table 6, *um*/*uh* ratios varied somewhat widely from region to region, especially among women, where the ratio ranged from 0.39 in the South Midland region, to nearly twice that (0.77) in the New England region. Despite this variation, a high level of gender differentiation was maintained across all regions. In all but one case, women's *um*/*uh* ratio exceeded men's by a factor of at least 2.5. The one exception was the Southern region, where the factor was 2.3.

Returning for a moment to the SDC, recall that in that corpus every conversation was between a female and a male. As a result, one might conjecture that the observed differences in filler usage across genders were a reflection not of the gender of the speaker, but of the gender of the interlocutor (or the combination of both interlocutors' genders). However, because the SWBC contains hundreds of mixed- and same-gender conversational dyads, we can investigate whether speaker or interlocutor gender appears to be more highly correlated with *um*/*uh* ratios among speakers in this corpus. Table 7 presents the *um* and *uh* data for each combination of the two coded genders:

Speaker Gender/ Hearer Gender	**Tokens of *um***	**Tokens of *uh***	**Combined**	***um*/*uh* ratio**
Women/Women	7,010	12,554	19,564	0.56
Women/Men	5,390	11,693	17,083	0.46
Men/Women	4,162	25,402	29,564	0.16
Men/Men	4,572	20,000	24,572	0.23

Table 7: Switchboard Corpus: *Um* v. *uh* by speaker and hearer gender.

[7] There was one additional region, "UNK", which is not depicted here. This region had only two speakers, both of whom were female. The "Mixed" category is included in Table 6 for completeness, despite the fact that (unlike the other categories displayed in the table) it does not represent speakers from a single dialect region.

In table 7, we see that speakers tended toward a higher *um*/*uh* ratio when speaking to members of the same gender than when speaking to members of the other gender: the *um*/*uh* ratio of women-to-women speech was 21.1% higher than that of women-to-men speech; and the difference in *um*/*uh* ratio between men-to-men speech and men-to-women speech was even greater, at 39.5%. Furthermore, these differences are highly statistically significant.[8] It is too soon to say whether this pattern is peculiar to the SWBC, or, if not, what would be a reasonable explanation of the results. In any case, these sizable differences don't seem quite so large in light of the differences between the *um*/*uh* ratios of male speakers and female speakers. Even when comparing women-to-men speech and men-to-men speech, we see that the *um*/*uh* ratio of the former exceeds that of the latter by a factor of two.

It should be noted that the *um*/*uh* ratios are far lower for the SWBC than those observed in the discussion of the SDC. In fact, we see that the average men's *um*/*uh* ratio of 0.61 in the SDC is higher than that of women in the SWBC, at 0.51. There may be several factors contributing to the higher *um*/*uh* ratios in the SDC. For one, the circumstances and purposes of the conversations comprising the two corpora are markedly distinct in kind: the speech in the SDC is entirely drawn from face-to-face conversations between people on dates, while the conversations in the SWBC were held between strangers speaking over the telephone about a predetermined topic of discussion. We may find linguistic variation across the two corpora based on these contextual differences alone. (Labov 1972, Rickford and McNair-Knox 1994, *inter alia*).

Yet another potential factor behind the cross-corpus differences in filler distribution is the fact that the SWBC predates the SDC by 15 years. The SWBC data provide some evidence of a linguistic change in progress: namely, the ascent of *um* with respect to *uh*. When the *um* and *uh* data in the SWBC are examined by age group, we find that the *um*/*uh* ratios tend to vary inversely with speaker age (Liberman 2010). Table 8 presents the results when speakers are grouped by their approximate age in 2011: less than 50,[9] 50–59, 60–69, and 70 or older. For both genders, we see the *um*/*uh* ratio drop with each incremental age cohort, so that the youngest women and men have *um*/*uh* ratios of 0.70 and 0.32, respectively, while the oldest women and men's *um*/*uh* ratios are much lower, at 0.26 and 0.09. Taking an apparent time perspective on the *um* and *uh* data in the SWBC (Gal 1978, Bailey 2002, *inter alia*), it appears that the popularity of *um* vis-à-vis *uh* may have been on the rise in American English at the time at which the SWBC was collected.

To be sure, even if this trend reflects a true change in progress, it would be premature at this stage to claim that the distinction in *um*/*uh* ratios between the SWBC and the SDC is born of such change, especially given the differences in conversational kind and context between the two corpora. Nonetheless, it is worth noting that the corpora were collected 15 years apart, and that the speakers in the SDC are, in general, far younger than those in the SWBC. In fact, all but one of the speakers in the SDC would fall into the age range of 28–40 at present, with an average age of approximately 33.[10] Having said that, we must leave a more thorough treatment of changes in the frequencies of *um* and *uh* for future research.[11]

[8] Women: $\chi^2 = 112.37$, 1 d.f., $p < 0.0001$; Men: $\chi^2 = 243.91$, 1 d.f., $p < 0.0001$. The test statistic for each gender was calculated based on comparing (i) the observed number of tokens of *um* for members of that gender speaking to the other gender and for members of that gender speaking to the same gender; and (ii) the expected number of tokens of *um* given the observed number of tokens of *uh* for each group.

[9] Only three speakers in the SWBC, all of whom were female, were born after 1971, so I did not include a "less than 40" age category in this analysis. Two of these speakers were born in 1972, and the third was born in 1975.

[10] Age range and average age values are based on the 107 speakers (out of 110) for whom age data are available.

[11] It should also be noted that the fillers *um* and *uh* taken together are far more prevalent in the SWBC: accounting for 2.96% of all words therein, compared with 1.10% in the SDC. Again, it is likely that this difference is the product of a multitude of factors. Note that there are also differences in the relative frequencies of other discourse particles across the two corpora. Perhaps most salient is the disparity in the frequency of the word *like*, which accounts for 2.26% of words in the SDC, but only 0.76% in the SWBC. Likewise, the word *so* is more than twice as frequent in the SDC (2.07% of words) as it is in the SWBC (0.89% of words).

Current Age	**Gender**	**Tokens of *um***	**Tokens of *uh***	***um*/*uh* ratio**
< 50	Women	3,232	4,611	0.70
	Men	3,262	10,141	0.32
50–59	Women	5,108	7,485	0.68
	Men	2,961	14,020	0.21
60–69	Women	2,360	5,663	0.42
	Men	1,367	9,032	0.15
70 +	Women	1,700	6,488	0.26
	Men	1,144	12,209	0.09

Table 8: Switchboard Corpus: *Um* v. *uh* by age and gender.[12]

2.3 Summary

In this section I provided evidence from two large, distinct corpora of English spoken in the United States that, on the aggregate, men's and women's distributions of *um* and *uh* are markedly different. In both the SDC and the SWBC, we observed that the average *um*/*uh* ratio among women was more than 2.5 times that of men. It was also demonstrated for both corpora that the differences were not merely the result of a small group of individuals. In addition, I showed that, among speakers in the SWBC, these aggregate gender differences persisted across dialect regions, and that while both men as a group and women as a group had higher *um*/*uh* ratios when talking to members of the same gender than when talking to members of the other gender, the men-to-men *um*/*uh* ratio was less than half of the women-to-men *um*/*uh* ratio. Finally, I presented preliminary evidence that *um* and *uh* have undergone a change in relative frequency over time, whereby *um/uh* ratios were (and may continue to be) on the rise.

3 What's the Meaning of All This?: Concluding, for Now

Having documented a large and robust difference between men and women's average *um*/*uh* ratios in both the Speed Dating Corpus and the Switchboard Corpus, I am left with the question of how to explain this difference. What are its sources and consequences? This question is beyond the scope of this paper. For now, I present a brief discussion of where we might look for answers.

Following, e.g., Eckert (1989), I take the position that an instance of gendered linguistic differentiation should not be assumed to be a direct function of gender *per se*. Rather, and more generally speaking, in analyzing linguistic variation along social lines, it is important to investigate how members of the social categories under consideration tend to stand in differing relation to their social world (Rickford 1986, Eckert 1989, *inter alia*), in part to better understand the stances, attitudes, and objectives that are relatively highly correlated with membership in those categories, while acknowledging variation within them. At the same time, one must consider the meanings and associations of the linguistic forms being examined, to the end of uncovering why speakers may be differentially inclined (consciously or not) toward using those forms. One's theories of these two aspects (i.e., the social and the semiotic) of the linguistic variation in question can thus be mutually informative: as one better understands the differential objectives and stances associated with speakers from a particular social group, one likewise better understands the semiotic value of the linguistic forms those speakers employ, and vice versa (see, e.g., Cameron et al. 1988).

Accordingly, in future research I intend to investigate how *um* and *uh* differ in what they communicate. I do not expect there to be an explanation as direct as, "*um* means 'female' and *uh* means 'male.'" Not only would such an analysis run counter to prior research on the relationship between social categories and social meaning (Ochs 1992, Podesva 2007, *inter alia*), it fails to account for other social patterns in filler usage presented here and elsewhere (Rayson et al. 1997,

[12] Ages are approximate, calculated as 2011 less year of birth.

Liberman 2010, Tottie 2011). It seems unlikely, for instance, that the filler usage of younger speakers in the SWBC, who used *um* at a higher rate than their older counterparts, was driven by an especially strong inclination to index "female."

The idea that *um* and *uh* serve distinct communicative functions is not without precedent. For instance, as at the outset of this paper, Clark and Fox Tree (2002) report that *um* generally signals longer pauses than *uh*. Moreover, Ward (2004) claims that /um/ tends to mean "thought-worthy" when used in what he calls "non-lexical conversational sounds." Do such findings hold across diverse corpora? If so, how can they help explain the distribution of *um* and *uh* along social lines?

Furthermore, how might an investigation of the meaning of *um* and *uh* shed light on the consistency among the age and gender findings of Tottie 2011, which draw upon the British National Corpus, and those presented herein, which are based on English spoken in the United States? It may be the case that both are instances of a more common trend whereby women lead in linguistic change (Labov 1990). But that alone does not explain why women would lead such a change or why the same change would occur across diverse populations and dialects. Is there something about the meanings of *um* and *uh* and women's relation to society in both cases that may be responsible for this change? Such questions must await future research.

Before concluding, I wish to emphasize that the meaning-based approach outlined above need not presuppose that all speakers of a particular social category are uniform in their position in and attitude toward the social landscape in which they live, nor do I advocate an approach that regards macro-level trends as conclusive. On the contrary, I hold that exceptions to generalizations may turn out to be just as informative as the original generalizations themselves. Finally, I stress also that this approach does not require of view of meaning as static. Rather, consistent with Eckert's (2008) notion of the indexical field, the meaning of a given linguistic form is expected to vary across time and individuals, so that with each use of that form, a speaker may at once draw upon and transform its meaning. *Um* and *uh*, in their ubiquity and the degree to which their use is socially stratified, provide a rich site for understanding these very dynamics.

References

Bailey, Guy. 2002. Real and apparent time. In *The Handbook of Language Variation and Change*, ed. J.K. Chambers, P. Trudgill, and N. Schilling-Estes, 312–331. Oxford: Blackwell.

Cameron, Deborah, Fiona McAlinden, and Kathy O'Leary. 1988. Lakoff in context: The social and linguistic function of tag questions. In *Women in Their Speech Communities: New Perspectives on Language and Sex*, ed. J. Coates and D. Cameron, 74–93. London: Longman.

Clark, Herbert H., and Jean E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speech. *Cognition*. 84:73–111.

Eckert, Penelope. 1989. The whole woman: Sex and gender differences in variation. *Language Variation and Change*. 1:245–267.

Eckert, Penelope. 2008. Variation and the indexical field. *Journal of Sociolinguistics*. 12:453–476.

Eckert, Penelope, and Sally McConnell-Ginet. 1992. Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology*. 21:461–90.

Fuller Medina, Nicté, and Joseph Roy. 2010. A variationist approach to sex: Statistics and the construction of social meaning. Paper presented at NWAV 39, The University of Texas at San Antonio.

Gal, Susan. 1978. Peasant men can't get wives: Language change and sex roles in a bilingual community. *Language in Society*. 7:1–16.

Jurafsky, Dan, Rajesh Ranganath, and Dan McFarland. 2009. Extracting social meaning: Identifying interactional style in spoken conversation. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 638–646.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William. 1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*. 2:205–254.

Liberman, Mark. 2010. The future of computational linguistics: Or, what would Antonio Zampolli do? Antonio Zampolli Prize Lecture presented at LREC 2010. Malta.

Macaulay, Ronald K.S. 1978. The myth of female superiority in language. *Journal of Child Language*. 5:353–363.

Ochs, Elinor. 1992. Indexing gender. In *Rethinking Context*, ed. A. Duranti and C. Goodwin, 335–359. Cambridge, U.K.: Cambridge University Press.

Podesva, Robert. 2007. Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics*. 11:478–504.

Rayson, Paul, Geoffrey Leech, and Mary Hodges. 1997. Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*. 2:133–152.

Rickford, John. 1986. The need for new approaches to social class analysis in sociolinguistics. *Language and Communication*. 6:215–221.

Rickford, John, and Faye McNair-Knox. 1994. Addressee- and topic-influenced style shift: A quantitative sociolinguistic study. In *Sociolinguistic Perspectives on Register*, ed. D. Biber and E. Finegan, 235–276. New York: Oxford University Press.

Switchboard-1 Release 2. 2010. Retrieved May 18, 2011, from http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97S62.

Tottie, Gunnel. 2011. *Uh* and *Um* as sociolinguistic markers in British English. *International Journal of Corpus Linguistics*. 16:173–197.

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

Department of Linguistics
Stanford University
Stanford, CA 94305-2150
eacton@stanford.edu