Uh and *Um* as sociolinguistic markers in British English*

Gunnel Tottie
University of Zurich

This study is based on the British National Corpus (BNC) and also takes data from the London-Lund Corpus (LLC) into account. It shows that the so-called filled pauses *er/uh* and *erm/um* are sociolinguistic markers that differentiate between registers of English and along gender, age and socio-economic class. Men, older people and educated speakers use more fillers than women, younger speakers and less educated speakers. Nasalization is used more often by women, younger speakers and more educated speakers. These sociolinguistic factors can probably partly explain the fact that the use of fillers is higher in the LLC and the context-governed sample of the BNC than in the demographic sample of the BNC. It is argued that a more positive view should be taken of fillers as planning signals, or *planners*, and that their functions should be submitted to careful discourse analytic study. Their recognition as words will facilitate such an undertaking.

Keywords: disfluency, filled pauses, hesitation markers, discourse markers, sociolinguistic markers, corpus linguistics

1. Introduction

In conversation, speakers of English tend to produce sounds that are not usually considered to be words, vocalizations consisting of schwa sounds, with or without nasalization, and with or without lengthening: [ə(:)] or [ə(:)m]. In written American English, they are usually transcribed as *uh* or *u(h)m*, and in British English, which is non-rhotic, as *er* or *erm*. Some examples from the British National Corpus (BNC) and the Santa Barbara Corpus (SBC) are given in (1) to (6):

(1) … yes, aha, so **er** Jim's been very busy (BNC)

(2) Oh yes Oh **erm**, but **er**, you know, wh — whether it'll be a good thing (BNC)

International Journal of Corpus Linguistics 16:2 (2011), 173–197. DOI 10.1075/ijcl.16.2.02tot
ISSN 1384–6655 / E-ISSN 1569–9811 © John Benjamins Publishing Company

(3) Yes of course, **er**, you know this thing we've been talking about (BNC)

(4) we shall go back, **erm** after Easter (BNC)

(5) [So every] — [2every2] **uh=**, horseshoe is made — custom-made for the horse then? (SBC)[1]

(6) **u=]m**, they're all different ages, (H) a=nd, .. **u=m**, they, … (H) .. you know, for the most part, they were probably very nervous… (SBC)

For simplicity's sake, I shall use the American-style transcriptions *uh* and *um*, which appear to be acceptable to speakers of British English as well as Americans.[2] I will use *uh+um* for aggregate uses of both variants.

The purpose of this paper is to describe and analyze the frequency of use of fillers in the English spoken by men and women, by speakers in different age groups, and by speakers from different socio-economic classes, as well as in different registers, and to show that they function as sociolinguistic markers in spoken British English. Uses of the nasalized *um* and non-nasalized *uh* variants as sociolinguistic markers will also be charted. In this endeavor, there are problems of terminology, and ultimately of linguistic and psycholinguistic theory, and it will therefore be necessary to first discuss some previous work.

2. Terminology

The first researchers that paid attention to the vocalizations under discussion were psycholinguists. In their seminal 1959 study Maclay & Osgood referred to them as 'filled pauses', including them among four different types of hesitation phenomena, the others being repeats, false starts, and silent pauses. The term 'filled pause' is also used by e.g. Goldman-Eisler (1961) and by the linguists Stenström (1990), Kjellmer (2003), and Gilquin (2008), but as pointed out by Kjellmer (2003: 190) it is an "anomalous term", as pauses are silent by definition. The psycholinguists Clark & Fox Tree (2002) refer to them as 'fillers', and so do Bortfeld et al. (2001) and Corley & Stewart (2008), who include them under the general heading "hesitation disfluencies". Corley et al. (2007) call them 'hesitations', and in his popularizing 2007 monograph, Erard refers to them simply as *ums*, under the heading "verbal blunders". Whatever the terms, these vocalizations are often treated as flaws or shortcomings — disfluencies — in an ideal world of fluent speech-production. Indeed, the term 'dis-' or 'dysfluency' is "not neutral", as pointed out by Boulton (2006). Others, like Kjellmer (2003) and Corley et al. (2007), stress the positive aspects of filler use (see further Section 7). Although in theory it could also be used about discourse markers like *you know* or *well*, I shall provisionally use the term 'filler' here to refer only to *uh* and *um*, but I will further discuss terminology in the concluding section, Section 7.

3. Are fillers words?

The status of fillers as words is also problematic. This is not just a matter of theoretical importance, but also involves practical considerations. Defining them as words will help legitimize their study as important items in discourse management, and if they are included in dictionaries, that will also facilitate their inclusion in curricula for English as a foreign or second language and help non-native speakers achieve more native-like competence in speech (cf. Stenström & Svartvik 1994).

As pointed out by Clark & Fox Tree (2002: 79) lexicographers have been slow to recognize the status of *uh* and *um* as words or word forms in spite of their frequency in spoken language: [ə:m] ranks as # 27 in the London-Lund Corpus (henceforth LLC), ranking higher than *think, as, so, no, with*, and [əm] comes in as # 75, much higher than *an, two, who* (cf. Svartvik et al. 1982: 43–49). Strangely enough, non-nasalized [ə(:)] is not listed among the top 100 — either because of transcription conventions or because it ranks lower.

However, there seems to be a trend towards greater recognition of *uh* and *um* among lexicographers. In the British *Collins Dictionary of the English Language* of 1979 only the non-nasalized variant [ə:], spelled *er,* is included, glossed as "*interj.* a sound made when hesitating in speech". Similarly, only *uh*, not the nasalized variant *um*, is included in the third edition of *The American Heritage Dictionary of the English Language* (1992) and glossed in the same way: "*interj.* used to express hesitation or uncertainty". However, the 2010 online edition of the same dictionary lists both *uh* and *um(m)*, classifying both as interjections and glossing them as shown in (7):[3]

(7)	*uh*	Used to express hesitation or uncertainty
	um(m)	Used to express doubt or uncertainty or to fill a pause when hesitating in speaking

Other online dictionaries vary: *Merriam-Webster* has *um* classified as an interjection, but not *uh, Encarta* has *um* "representing hesitation in speech" but *uh* is (surprisingly) called a "grunting exclamation expressing surprise". *Google Dictionary* has *uh* and *um* and, also surprisingly, the spelling *er*, glossing them as interjections used to express hesitation. The online *Oxford English Dictionary* includes *er*, glossed as in (8), without any part-of-speech classification:

(8) *er* Used to express the inarticulate sound or murmur made by a hesitant speaker.

The nasalized variant is not represented as *erm*, but — inconsistently — as *um*, and is classified as an interjection:

(9) *um* 1. Used to indicate hesitating or inarticulate utterance on the part of a speaker.
2. Used to indicate hesitation or doubt in replying to another.

It is interesting that dictionaries do not as a rule give straightforward meanings of *uh* and *um* but prefer the term "used to express/indicate" — it is thus a kind of procedural meaning that is indicated, and direct glosses are avoided.

Among linguists and psycholinguists, there has also been some discussion concerning the status of fillers as words. Clark & Fox Tree (2002) and Shillcock et al. (2001) argue that they are words, and O'Connell & Kowal (2005: 573) find them "worthy of legitimation as words, but requir[ing] much more empirical research to specify their meaning", but others disagree. Kjellmer (2003: 190) holds that they are not words, because speakers do not include them when asked to repeat or clarify an utterance containing one. Corley & Stewart (2008: 589) also contend that they are not words in the conventional sense on the grounds that "there is little evidence to suggest that they are intentionally produced". However, both of these arguments can be countered with the fact that even bona fide words like *well* or *you know* used in similar functions are rarely produced intentionally, and they would probably not be included if someone was asked to repeat an utterance containing one of them.[4] Notice also that like English, other languages use conventional words in the same functions, e.g. Latin American Spanish *este* or German *also* (cf. Quinting 1971). So although fillers may not be prototypical words, they are at least "marginal words" as suggested in Du Bois et al. (1992).

4. What can corpus linguistics contribute?

If we want to know how fillers are really used and how they function in discourse, we clearly need to study corpora consisting of naturally occurring speech. However, most work on fillers has been carried out by psychologists or psycholinguists using experimental data or material constructed for the purpose of experiments. The great exception as regards data is a seminal study by Clark & Fox Tree (2002). They use 170,000 words from the LLC, consisting of educated British English recorded in the 1960's and 70's (see Svartvik & Quirk 1980) and also present results based on other corpora, the most important being the Switchboard Corpus (SWB), consisting of 2.7 million words of American English telephone conversation between individuals who did not know each other. Shriberg (1994), a psychologically oriented phonetician, also analyzed SWB and other corpora for a dissertation on disfluencies.[5]

Few professed corpus linguists have studied fillers systematically and subjected them to quantitative analysis, but a few pioneering works exist. Stenström (1990) based her study on ten texts (50,000 words) from the LLC and compared "filled pauses" with "silent pauses", finding that vocalizations were much less frequent than silent pauses. Kjellmer (2003) based his study on 57.4 million words from the Cobuild corpus, including both spoken and written data. His work presents quantitative data concerning items that *er* and *erm* co-occur with, and an excellent but non-quantified discussion of their various discourse functions. Gilquin (2008) discusses the use of "filled pauses" in the language of learners of English as a foreign language, comparing it with their use by native speakers, taking her native speaker data from the Louvain Corpus of Native English Conversation (LOCNEC: 125,226 words). Svalduz (2006) also provides quantitative data: he compares the use of fillers in the spoken component of the British National Corpus (BNC-S) and the SBC. Finally, a sociolinguistic study of vocabulary frequencies based on BNC-DEM (see below) by Rayson et al. (1997) includes both *er* and *erm* among a large number of other lexical items. The authors show that *er* is significantly more frequent in male than in female speech, and also more frequent in the speech of people over 35 than among younger people. *Erm* is shown to be more frequent in the speech of the highest social groups than elsewhere, but the use of the two items is not linked or further discussed.

I will take the results of the above studies as my point of departure and first give an overview of the properties of the above-mentioned corpora in Table 1.

Table 1. Overview of corpora used for research on fillers

Variety	Corpus	No. of words used	Recording date	Characteristics
British	LLC	c. 170,000	1960's	Informal conversation, academics
British	BNC-DEM	4,233,962	c. 1990	Informal conversation, varying social strata
British	BNC-CG	6,175,896	c. 1990	Domain-specific: business, educational/informative, leisure, public or institutional
British	LOCNEC	125,226	1995–1996	Interviews with students
American	SBC	68,000	1980's	Informal conversation, varying social strata
American	SWB	2,700,000	prior to 1992	Telephone conversation

Data from BNC-S are divided into two categories, coming from two different types of spoken language. The context-governed part (BNC-CG) consists of recordings from four so-called "domains": business, education, leisure, and public/institutional, with further subdivisions. BNC-CG has been characterized as more "formal" — something that needs to be further discussed. The demographic part (BNC-DEM) consists of impromptu speech in informal settings. To a large extent, speakers are also classified according to sociolinguistic criteria: age, gender and socio-economic class (see further Hoffmann et al. 2008, Chapter 3).

The distribution of fillers in the different corpora is shown in Table 2. As corpus sizes vary widely, from 68,000 words to over 6 million, results here and be-

Table 2. Absolute and relative frequency per 100,000 words of fillers in LLC, BNC-DEM, BNC-CG, LOCNEC, SBC and SWB[7]

Corpora, No. of words	*uh*	%	*um*	%	Totals	n/100K
LLC	2,111	54%	1,793	46%	3,904	2,297
BNC-DEM	21,345	56%	16,605	44%	37,950	896
BNC-CG	67,009	59%	45,747	41%	112,756	1,823
LOCNEC	1,047	35%	1,910	65%	2,957	2,361
SBC	216	47%	240	53%	456	671
SWB	67,065	84%	12,558	16%	79,623	2,949

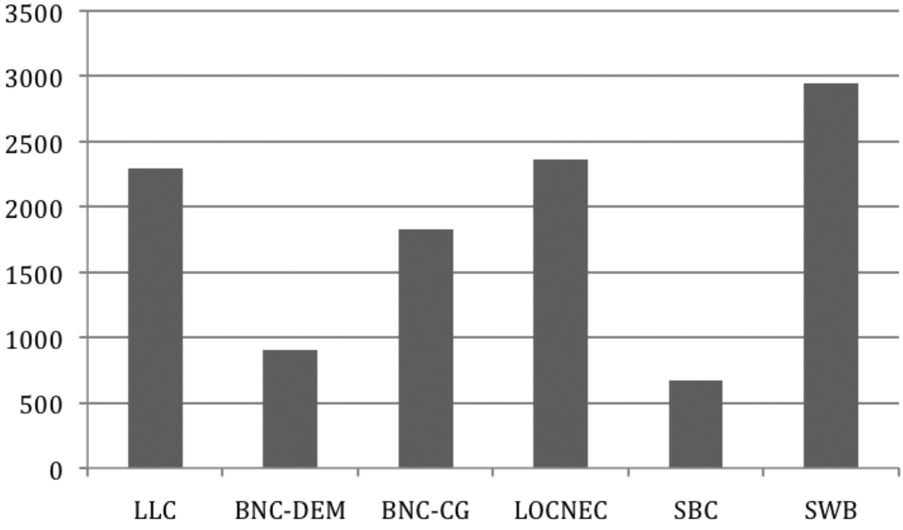

Figure 1. Frequency of *uh+um* in six different corpora of spoken English (per 100,000 words)

low are given as frequencies per 100,000 words rather than the more common per-million-word measure, to avoid extrapolation and unwieldy high numbers. Proportions of *uh* and *um* are given in this table but lengthening is not taken into account, as it is not available for most of the corpus material.[6]

The totals of *uh+um* are displayed graphically in Figure 1.

Figure 1 shows great differences between the different corpora, but no obvious divide between British and American English. The highest frequency of fillers is found in the American SWB conversations — 2,949/100K, followed by the British LOCNEC and LLC with 2,361 and 2,297 instances, respectively. With a frequency of 1,823/100K, BNC-CG comes much closer to these frequencies than BNC-DEM, which has only 896 fillers per 100K words. SBC has the lowest frequency of all: only 671 per 100K words. The question now is: can we account for this variation and the differences between the corpora?

The SWB data will not be included in the discussion, as it has been established by Shriberg (1994) and others that telephone conversations have more vocalizations than face-to-face spoken interaction. LOCNEC is also less comparable to the other corpora as it consists of elicited answers to questions. In this paper I will focus on British English, leaving the SBC for future consideration, and concentrate on the differences between BNC-DEM and BNC-CG. I shall examine these two sub-corpora for sociolinguistic factors influencing the use of *uh* and *um* in Section 5 and discuss their possible importance for frequency differences in Section 6.1.[8] The differences between BNC-DEM and the LLC will be considered in Section 6.2. I will summarize the results and discuss theoretical implications as well as prospects for future research in Section 7.

5. Fillers as sociolinguistic markers in BNC-DEM and BNC-CG

The difference between the two parts of BNC-S has often been characterized as one of formality — in the words of Hoffmann et al. (2008: 34), the texts of BNC-CG "tend to contain more formal language use". It is thus surprising that the frequency of fillers is higher in BNC-CG than in BNC-DEM, as they have often been associated with informality (cf. Clark & Fox Tree 2002: 98f). It certainly is a fact that very formal speeches such as presidential addresses, radio and TV speeches are usually devoid of fillers, but formality is a fuzzy concept and may not be a determining factor here. It therefore seemed interesting to consider whether sociolinguistic factors might contribute to explaining the differences between BNC-DEM and BNC-CG: gender, age and socio-economic class. BNC-S is annotated for all of these, but it is important to keep in mind that the annotation is not complete: BNC-DEM and BNC-CG are not annotated to the same extent for the three factors, which means that the sizes of the subsamples with different kinds of annotation differ. The tables below are thus based on different — and lower — numbers than those given for the whole spoken component in Table 2.

Another caveat is necessary. According to Clark & Fox Tree (2002) *uh* and *um* tend to occur with pauses of different lengths, *um* collocating with longer silent pauses than *uh*. It also seemed interesting to try to find out if nasalization was conditioned by the sociolinguistic factors investigated for this paper. Sound files are not currently available for the BNC, but according to Hoffmann et al. (2008: 38) "[i]n the case of filled pauses, a fairly high level of consistency was achieved because transcribers were explicitly instructed to use *er* or *erm* to capture the wide range of potential variants". Yet, as we cannot check how consistent the transcribers (using *er* and *erm*) have been in individual cases, results must be regarded as preliminary.

5.1 Gender

Earlier research based on telephone conversation (Shriberg 1994), experimental data (Bortfeld et al. 2001) and to some extent on the BNC (Rayson et al. 1997) has shown that women and men tend to differ in their uses of fillers, with men using more and women fewer. As BNC-DEM and BNC-CG are each other's opposites as regards gender distribution, with BNC-DEM having two-thirds female speakers and BNC-CG two-thirds male speakers, it seemed conceivable that this could be an explanatory factor and that it would be worthwhile to study frequencies in more typical and natural data than Shriberg or Bortfeld et al. had done and treat *uh+um* as one variable rather than as individual words as done by Rayson et al. Table 3 shows the distribution of fillers in BNC-DEM and BNC-CG according to gender.

In Table 3, column 2 shows the number of words classified for the factors male and female, totaling 3,718,438 words from BNC-DEM and 4,522,069 from BNC-CG, and thus less than the totals for each sub-corpus but still very sizable samples. The numbers of words examined for each gender are also indicated in this column, and in column 3, those numbers have been recalculated as percentages of the sub-corpora. We see that women dominate in BNC-DEM, with 61%, and that men dominate even more sharply in BNC-CG, with 77%. Column 4 lists the variants *uh* and *um* as well as the totals of *uh+um* in boldface. Column 5 shows the number of speakers who actually used the fillers — the dispersion — and column 6 shows dispersion as percentages, ranging from 59% to 75%, indicating that there are both "ummers" and "um-avoiders", to use Christenfeld's (1995: 171) terms. Column 7 shows the total number of instances recorded, "hits", and column 8 displays the percentages of fillers produced by male and female speakers. Although men make

Table 3. Overall distribution of fillers by gender in BNC-DEM and BNC-CG

1	2	3	4	5	6	7	8	9	10
Gender	**Total no. words**	**M/F %**	**Filler**	**Dispersion**	**Disp. %**	**No. hits**	**M/F % hits**	**Freq/ 100K**	**% um**
DEMOGRAPHIC SAMPLE									
Male	1,454,344	39%	uh	377/509	74%	9,415	44%	647	
			um	333/509	65%	5,153		354	35%
			uh+um	–	–	14,568		**1,001**	
Female	2,264,094	61%	uh	418/559	75%	9,337	56%	412	
			um	415/559	74%	9,069		401	49%
			uh+um	–	–	18,406		**813**	
Totals	**3,718,438**	**100%**				**32,974**		**886**	
CONTEXT-GOVERNED SAMPLE									
Male	3,495,594	77%	uh	1,456/1,939	75%	40,534	83%	1,160	
			um	1,182/1,939	61%	25,787		737	39%
			uh+um	–	–	66,321		**1,897**	
Female	1,026,475	23%	uh	492/801	61%	7,084	17%	690	
			um	474/801	59%	6,727		655	49%
			uh+um	–	–	13,811		**1,345**	
Totals	**4,522,069**	**100%**				**80,132**		**1,772**	

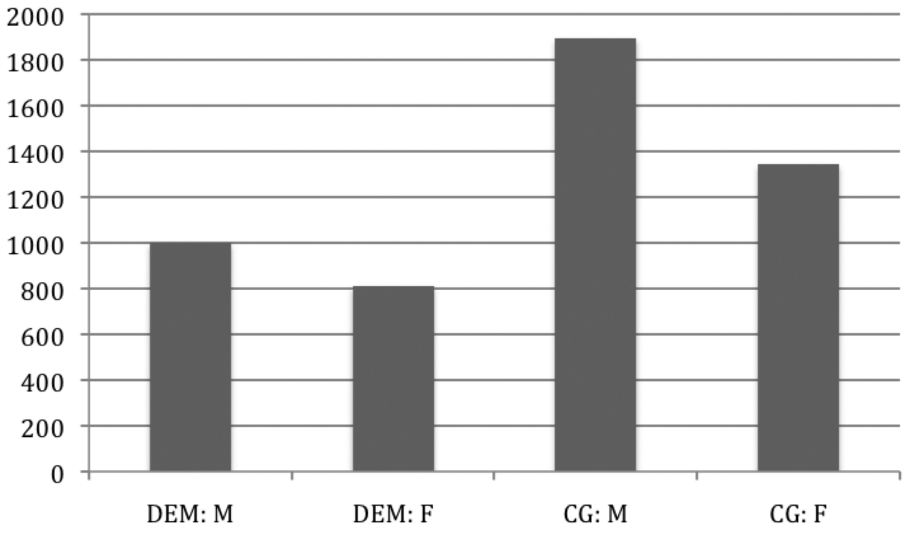

Figure 2. Frequency per 100,000 words of *uh+um* according to gender in BNC-DEM and BNC-CG

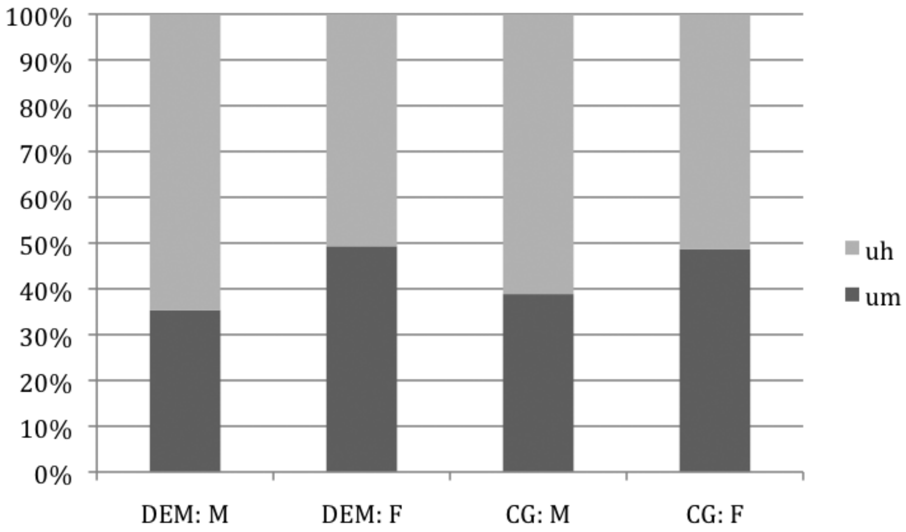

Figure 3. Proportions of *uh* and *um* according to gender in BNC-DEM and BNC-CG

up 39% of the BNC-DEM sample, they produce 44% of the fillers, and in BNC-CG, they account for 77% of the sample but 83% of the fillers, thus more than their expected share, but not spectacularly so. Column 9 shows the frequency of fillers per 100,000 words: 886 in BNC-DEM and 1,772 in BNC-CG, and finally, column 10 indicates the proportions of the nasalized filler *um*. Figure 2 gives a graphic display of the distribution of the fillers *uh+um* across male and female speakers in BNC-DEM and BNC-CG.

We see that in both BNC-DEM and BNC-CG, men display a higher frequency of fillers than women, 1,001/100K vs. 813/100K in BNC-DEM and 1,897/100K vs. 1,345/100K in BNC-CG (both in BNC-DEM and in BNC-CG, the differences are significant, $p < .0001$).[9] But both genders increase their rates of fillers in BNC-CG; in fact, women have a higher total frequency of fillers in BNC-CG than men have in BNC-DEM.

It is also clear from column 10 in Table 3 and from Figure 3 above that the use of nasalized and non-nasalized fillers differs between male and female speakers. In both BNC-DEM and BNC-CG, women have higher proportions of *um*: 49% of all fillers compared with 35% for male speakers in BNC-DEM, and 49% vs. 39% for male speakers in BNC-CG. The difference between the genders is 14 percentage points in BNC-DEM and less in BNC-CG: 10 percentage points. Both in BNC-DEM and in BNC-CG, the differences are significant ($p < .0001$).[10]

These results clearly support earlier research concerning male/female frequency differences as regards the use of fillers. Gender is a powerful sociolinguistic variable: men use more fillers than women in impromptu conversation as well as in more circumscribed contexts, but women use a higher proportion of nasalized fillers than men.

5.2 Age

Another factor will be examined next: speaker age. Bortfeld et al. (2001) showed in an experiment that older speakers (aged 63–72) tended to use more fillers than younger speakers, and Rayson et al. (1997) showed that speakers over 35 used more instances of *er (uh)*. The spoken component of the BNC is to a large extent annotated for age, with the material divided into six age cohorts: 0–14, 15–24, 25–34, 35–44, 45–59, and 60+.[11] A much larger proportion of BNC-DEM than of BNC-CG is annotated for age: 86% of BNC-DEM (3,657,428 out of a total of 4,233,962 words) vs. only 37% of BNC-CG (2,294,269 out of a total of 6,175,896 words), but these numbers are still large enough to form representative samples. A worse problem is the fact that the two youngest cohorts, 0–14 and 15–24, are under-represented in both corpora. BNC-DEM still has fairly sizeable samples, accounting for 10% and 14% of the annotated sample, respectively, but in BNC-CG the proportions for these cohorts dwindle to 4% for the 15–24 age group and 1% for the 0–14 age group, which is represented by less than 30,000 words (cf. Figure 4).

The use of fillers by different age groups in BNC-DEM and BNC-CG is accounted for in Table 4, which is arranged in the same way as Table 3, but for clarity information concerning nasalization is given in a separate table, Table 5 below.

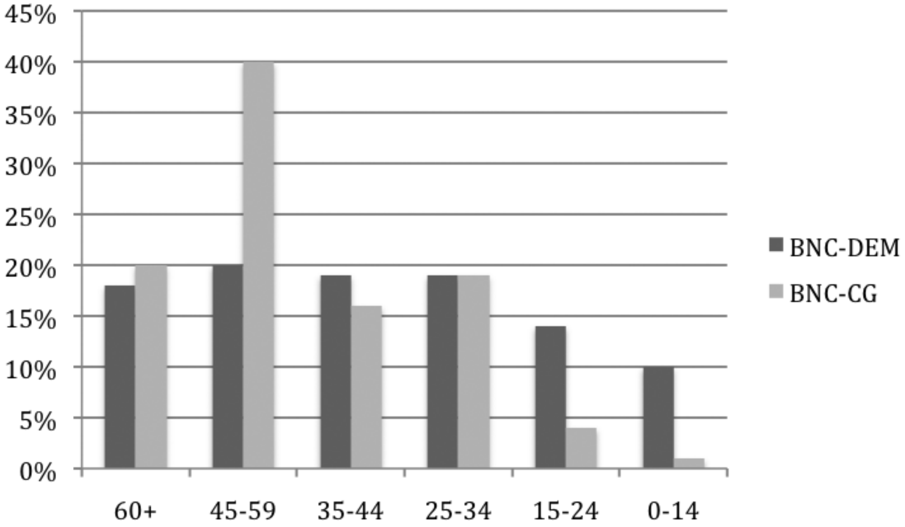

Figure 4. Proportions of age cohorts in BNC-DEM and BNC-CG

Table 4. Overall distribution of fillers by age in BNC-DEM and BNC-CG

1	2	3	4	5	6	7	8	9
Age	**Total no. words**	**Age %**	**Filler**	**Disper-sion**	**Disp. %**	**No. hits**	**Age % hits**	**Freq/ 100K**
DEMOGRAPHIC SAMPLE								
60+	671,392	18%	uh	122/142	86%	5,418	26%	807
			um	107/142	75%	2,913		434
			uh+um	–	–	8,331		**1,241**
45–59	733,141	20%	uh	116/153	76%	4,240	21%	578
			um	113/153	74%	2,534		346
			uh+um	–	–	6,774		**924**
35–44	705,882	19%	uh	115/147	78%	2,857	16%	405
			um	101/147	69%	2,345		332
			uh+um	–	–	5,202		**737**
25–34	690,721	19%	uh	128/163	79%	2,745	16%	397
			um	114/163	70%	2,571		372
			uh+um	–	–	5,316		**769**
15–24	500,619	14%	uh	145/211	69%	2,001	12%	400
			um	147/211	70%	2,012		402
			uh+um	–	–	4,013		**802**
0–14	355,673	10%	uh	135/201	67%	1,274	8%	358
			um	135/201	67%	1,468		413
			uh+um	–	–	2,742		**771**
Totals	**3,657,428**	**100%**				**32,378**		**885**
CONTEXT-GOVERNED								
60+	466,041	20%	uh	142/176	81%	9,720	26%	2,086
			um	86/176	49%	2,305		495
			uh+um	–	–	12,025		**2,581**
45–59	905,223	40%	uh	244/283	86%	8,121	32%	897
			um	227/283	80%	6,497		718
			uh+um	–	–	14,618		**1,615**
35–44	369,867	16%	uh	164/188	87%	4,426	16%	1,197
			um	136/188	72%	2,938		794
			uh+um	–	–	7,364		**1,991**

Table 4. (*continued*)

1	2	3	4	5	6	7	8	9
Age	**Total no. words**	**Age %**	**Filler**	**Disper-sion**	**Disp. %**	**No. hits**	**Age % hits**	**Freq/ 100K**
25–34	429,796	19%	uh	156/188	83%	4,151	19%	966
			um	147/188	71%	4,555		1,060
			uh+um	–	–	8,706		**2,026**
15–24	93,781	4%	uh	68/91	75%	928	5%	990
			um	70/91	77%	1,193		1,272
			uh+um	–	–	2,121		**2,262**
0–14	29,561	1%	uh	28/57	49%	262	1%	886
			um	26/57	46%	299		1,012
			uh+um	–	–	561		**1,898**
Totals	**2,294,269**	**100%**				**45,395**		**1,978**

The data are summarized in Figure 5, which shows the frequency of fillers in different age groups in BNC-DEM and BNC-CG.

In both sub-corpora the oldest speakers have the highest frequencies of *uh+um*. If we compare the 60+ age groups with all speakers under 60, there are significant differences in both BNC-DEM and BNC-CG ($p < .0001$ in both subcor-

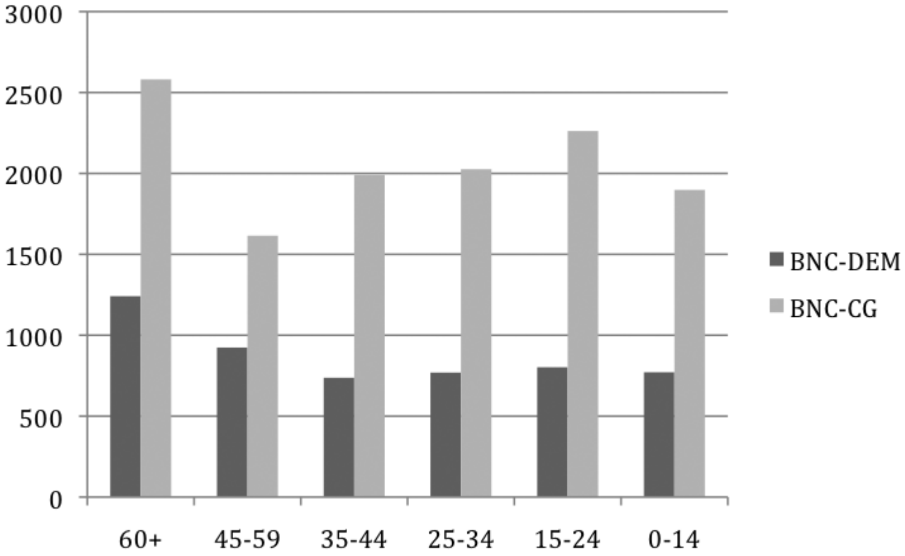

Figure 5. Frequency per 100,000 words of *uh+um* across age cohorts in BNC-DEM and BNC-CG

Table 5. The use of *uh* and *um* across age cohorts in BNC-DEM and BNC-CG

BNC-DEM					BNC-CG			
Age	**uh**	**um**	**% um**	**uh+um**	**uh**	**um**	**% um**	**uh+um**
60+	807	434	**35%**	1,241	2,086	495	**19%**	2,581
45–59	578	346	**37%**	924	897	718	**34%**	1,615
35–44	405	332	**45%**	737	1,197	794	**40%**	1,991
25–34	397	372	**48%**	769	966	1,060	**52%**	2,026
15–24	400	402	**50%**	802	990	1,272	**56%**	2,262
0–14	358	413	**54%**	771	886	1,012	**53%**	1,898

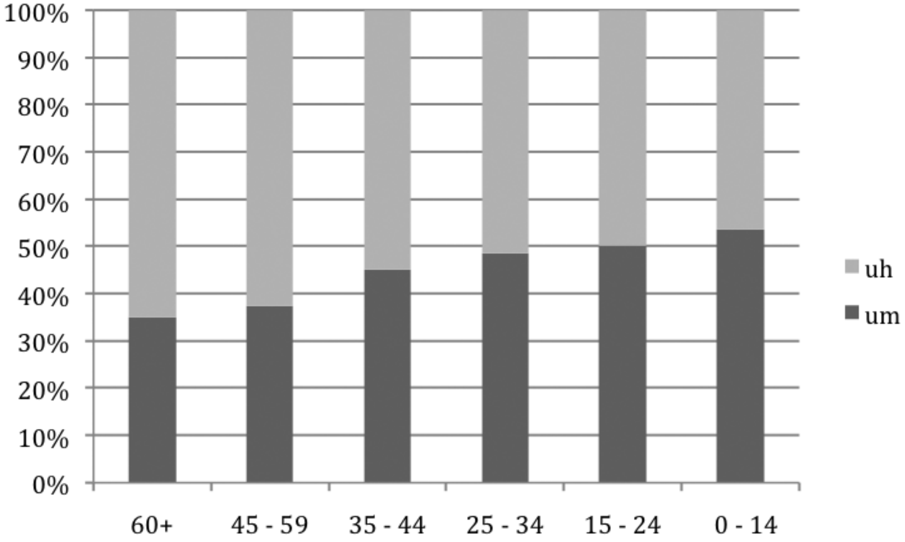

Figure 6a. Proportions of *uh* and *um* in different age cohorts in BNC-DEM

pora).[12] Otherwise there is no consistent tendency towards increase or decrease among the age groups. In BNC-DEM the lowest frequency, 737/100K, is recorded for the 35–44 cohort, but in BNC-CG the 45–59 cohort has the lowest frequency, 1,615/100K. The results thus do not give any clear indication either of change over time or age-grading but do square well with the findings of Bortfeld et al. (2001) that the oldest speakers use more fillers than younger ones.

However, turning next to nasalization, we can discern what seems to be a clear trend in both BNC-DEM and BNC-CG, viz. an increased use of the nasalized variant *um* in the younger cohorts, as appears from Table 5 and Figures 6a and 6b.

There is a very steady progression in both sub-corpora: the proportion of *um* goes up from 35% among the over-sixties to 54% among the under-fourteens in

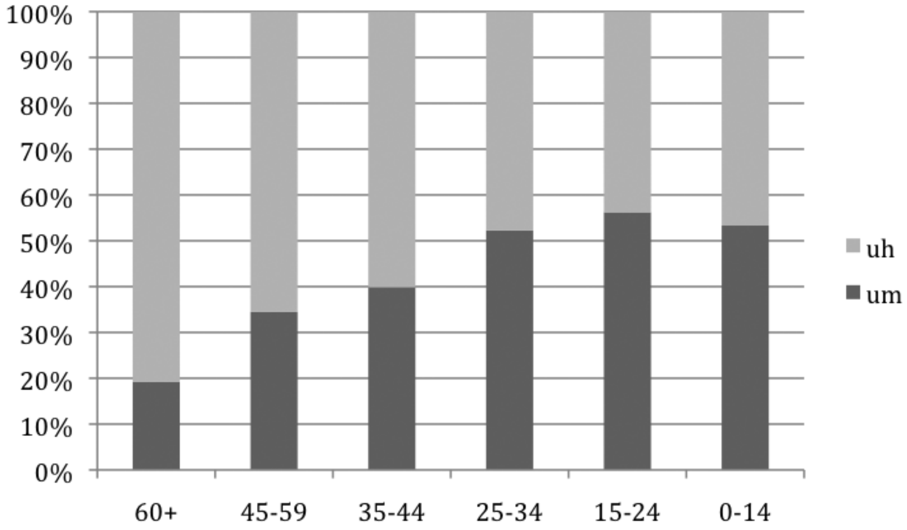

Figure 6b. Proportions of *uh* and *um* in different age cohorts in BNC-CG

BNC-DEM, and from 19% among the oldest speakers to over 50% in the youngest cohorts in BNC-CG. In BNC-CG the rise is steady from 60+ to 15–24 (which has 56% *um*); the fact that it then goes down by 3 percentage points to 53% among the under-fourteens could be due to the smallness of the sample of the youngest speakers.

Like gender, age is thus a sociolinguistic factor of importance, not so much as regards the overall use of fillers, but definitely for the choice of filler variant: the proportion of nasalization increases steadily as speakers' ages go down.

5.3 Fillers and socio-economic factors

A third sociolinguistic factor, which is at least in part testable based on available BNC data, is socio-economic class, as the BNC is annotated for four categories as listed below (see Hoffmann et al. 2008: 35):

- AB (top or middle management, administrative or professional)
- C1 (junior management)
- C2 (skilled manual)
- DE (semi-skilled or unskilled)

However, social class is documented for only about a quarter of the spoken material, and almost exclusively for the demographically sampled sub-corpus. In BNC-CG, only 138,296 words are annotated for social class, with no data from groups C1 and C2, and only 27% from DE, yielding a very unsatisfactory sample (cf. Hoff-

Table 6. Overall distribution of fillers across socio-economic (SEC) groups in BNC-DEM

1	2	3	4	5	6	7	8	9	10
SEC group	**Total no. words**	**SEC grp %**	**Filler**	**Disper-sion**	**Disp. %**	**No. hits**	**SEC grp % hits**	**Freq/ 100K**	**% um**
AB	716,328	27%	uh	73/86	85%	3,526	30%	492	
			um	72/86	84%	3,473		485	48%
			uh+um	–	–	6,729		**977**	
C1	782,234	30%	uh	104/114	91%	3,583	30%	458	
			um	100/114	88%	3,160		404	47%
			uh+um	–	–	6,743		**862**	
C2	719,884	27%	uh	83/99	84%	3,526	25%	452	
			um	70/99	71%	2,006		279	36%
			uh+um	–	–	5,532		**731**	
DE	414,066	16%	uh	52/59	88%	2,220	15%	536	
			um	47/57	82%	1,020		246	31%
			uh+um	–	–	3,240		**782**	
Totals	**2,632,512**	**100%**				**22,244**		**845**	

mann et al. 2008: 36). The results presented in Table 6 — arranged according to the same principles as Tables 3 and 4 — are therefore based entirely on BNC-DEM, which offers 2,632,512 words of annotated text. As appears from the table, there are robust samples from all four socio-economic groups in BNC-DEM, each accounting for between 16% and 30% of the sample.

Figure 7 provides a graphic illustration of the use of *uh+um* by the four socio-economic categories in BNC-DEM. Members of the highest socio-economic class, AB, have the highest frequency of fillers, 977/100K, followed by junior management, C1, with 862/100K, and skilled manual workers, C2, who have only 731/100K. Semi-skilled and unskilled workers, DE, go counter to the trend and have 782/100K, thus somewhat more than the next “higher” class, but they are still well below C1. The differences are statistically significant ($p < .0001$).[13]

Socio-economic class also proves to be an important factor for the choice of nasalized or non-nasalized variant, as appears from column 10 in Table 6 and Figure 8. AB and C1 speakers have the highest proportions of *um* — 48% and 47% of

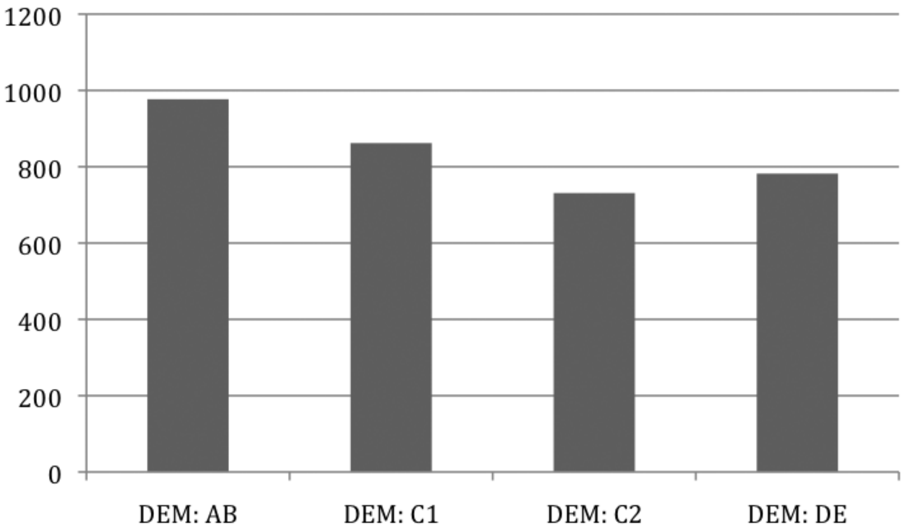

Figure 7. Frequency per 100,000 words of *uh+um* across socio-economic groups in BNC-DEM

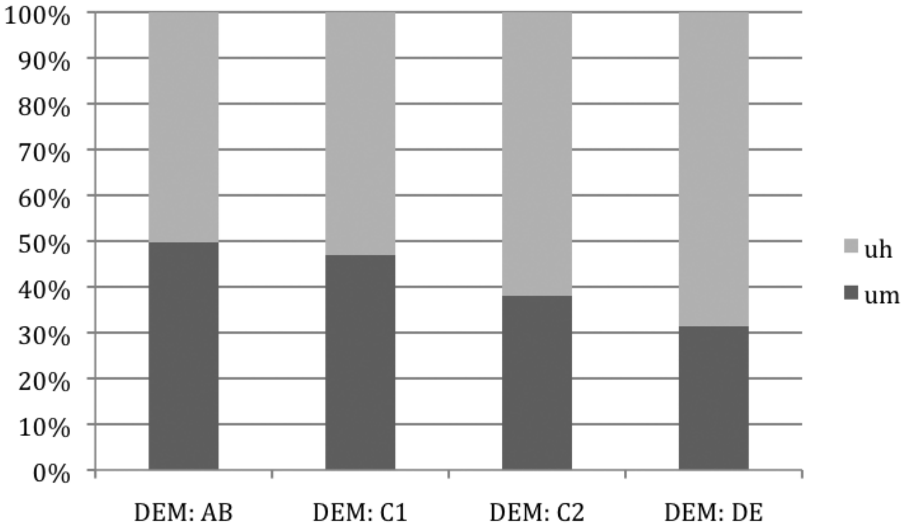

Figure 8. Proportions of *uh* and *um* in socio-economic groups in BNC-DEM

the totals, respectively. There is then an 11 percentage-point gap between C1 and C2 speakers, who use 36% *um*, while DE speakers use even fewer: only 31%. The big divide is thus between management on the one hand and skilled, semi-skilled or unskilled manual workers on the other. The differences are statistically significant ($p < .0001$).[14]

So, to conclude this section, although the overall favorite filler is *uh* in both BNC-DEM and BNC-CG, with 56% and 59% of the totals, respectively, we have seen that the nasalized filler *um* is used more by women than by men, more by younger speakers than by older ones, and more by speakers of the managerial classes than by those of the manual-worker classes. But it is the aggregate data that we turn to in order to try to explain the differences between corpora and sub-corpora.

6. Why are there more fillers in some corpora than in others?

6.1 BNC-DEM vs. BNC-CG

Turning first to the BNC, we can now be certain that men use more fillers than women: as shown in Table 3, women have only 813/100K compared with 1,001/100K for men in BNC-DEM, and 1,345/100K compared with 1,897/100K for men in BNC-CG. This means that women have 81% of the male ratio in BNC-DEM and 71% in BNC-CG. Although the higher proportion of male speakers in BNC-CG (c. two-thirds, to be compared with one-third in BNC-DEM) will account for some of the difference in fillers, it will not suffice to entirely explain the much higher use of fillers in BNC-CG — notice that both men and women increase their use of fillers in BNC-CG.[15] Nor does age help explain the difference, as the oldest cohorts, where fillers are most frequent, account for similar proportions in BNC-DEM and BNC-CG (18% vs. 20%) — assuming of course that the reduced samples with annotation for age mirror the proportions of the entire sub-corpora. Socio-economic class is a better bet — we have seen that speakers from the managerial classes exhibit higher frequencies of fillers in BNC-DEM (977 and 862/100K for AB and C1 vs. 731 and 782/100K for C2 and DE, thus about 25% more), and although annotation according to socio-economic class is largely missing for BNC-CG, it seems reasonable to assume that the proportion of AB speakers is high in that part of the corpus, which would account for a proportionally larger number of fillers in BNC-CG.

But in order to be certain of the importance of individual factors, we would need to check their interaction by means of a multivariate analysis using Varbrul or other statistical tools (see e.g. Tagliamonte 2006, Gries 2009). This is beyond the scope of the present study, and this work must thus be regarded as a preliminary to such an undertaking.

6.2 Comparing BNC-DEM and LLC

As shown in Table 2, the difference in frequency of fillers between BNC-DEM and the LLC is even larger than that between BNC-DEM and BNC-CG: 896/100K in BNC-DEM vs. 2,297/100K in the LLC, significant at $p < .0001$.[16] LLC speakers thus use two and a half times as many fillers as BNC-DEM speakers, even though both corpora contain supposedly similar language use — informal impromptu conversation, mostly between speakers who know each other, in the same variety of the language, British English. One factor seemed to have potential for explaining the different frequency of fillers: age (and possibly language change). The age group that stands out in the BNC as displaying a higher frequency of *uh+um* is the 60+ cohort, both in BNC-DEM and BNC-CG. The LLC data were collected some thirty years earlier than those of the BNC, and it seemed conceivable that there could have been a change over time in the use of fillers, perhaps in favor of discourse markers like *well, you know* or *like* (see e.g. Andersen 2001, Chapter 5). But although there is a fairly steady decrease in BNC-DEM as we move towards the younger cohorts, the open-endedness of the 60+ age group creates a problem. It is known from psychological research (e.g. Bortfeld et al. 2001) that the frequency of fillers rises with age, most likely because of a slowing down of cognitive functions that necessitates more time to retrieve words. We cannot therefore assume that the lower ratio of fillers in BNC-DEM is a result of language change.

The gender composition of the LLC is not discussed by Clark & Fox Tree (2002), but Oreström (1983) shows that a sample from the same part of the LLC that was investigated by Clark & Fox Tree comprises about one-third female and two-thirds male speech, and this is corroborated by data from the Diachronic Corpus of Present-Day Spoken English, (DCPSE).[17] Although it seems certain that the high frequency of fillers recorded for the LLC is to some extent due to the strong male presence, it is also a fact that women display a very high frequency of fillers in this corpus (cf. endnote 17, Table i).

Something that may also account for the high frequency of fillers in the LLC is the third, socio-economic factor. It is clear from Svartvik & Quirk (1980: 26–31) that virtually all LLC speakers are academics or have some form of higher education, so although they would not be labeled "managerial", they would fit in at the top of the socio-economic scale, the AB stratum. The evidence from both the LLC and BNC suggests that this could be the most important sociolinguistic factor influencing the use of *uh+um*, but more research based on other corpora is necessary to validate this conclusion.

7. Conclusions and prospects

As far as can be ascertained on the basis of the British National Corpus without a multifactorial analysis, I have shown that the filler *uh/um* — which can be regarded as variants of a single variable — constitutes a sociolinguistic marker, and that the aggregate use of the variants is conditioned by three factors: speaker gender, speaker socio-economic status, and to a lesser degree, speaker age. Men use more fillers than women, speakers from higher socio-economic and more well-educated strata use more fillers than speakers from lower strata, and so do the oldest speakers, aged over 60. However, the evidence from younger cohorts is not conclusive; there is probably age-grading but not linguistic change.

The choice of variant is even more sensitive to sociolinguistic factors: although the non-nasalized *uh* is more frequent overall than nasalized *um* with over 50% of all instances, women show a higher ratio of *um* than men, and there is a steady increase from older to younger age groups in both BNC-DEM and BNC-CG, so that it is conceivable that there is linguistic change going on. Finally, there is a clear differentiation in the use of *um* between socio-economic strata: AB speakers have almost 50% nasalization, and the proportion goes down steadily to just over 30% in DE. An interesting question for future research is how these findings square with the findings of Clark & Fox Tree (2002) and Fox Tree (2002) that nasalized fillers are linked to longer pauses.

Differences in the frequency of fillers between BNC-DEM and BNC-CG were shown to be likely to be due in part to their having different proportions of male and female speakers: the two-thirds majority of men in BNC-CG is likely to be one reason for the higher frequency of fillers in that corpus, and similarly, the fact that men are a minority in BNC-DEM can probably be linked to the lower frequency of fillers in that corpus. Age does not seem to be a factor, but although data on the socio-economic composition of BNC-CG are almost entirely missing, it seems highly likely that the majority of speakers would belong to the highest socio-economic groups and therefore be likely to have a high frequency of fillers. However, these factors do not suffice to explain the fact that BNC-CG has twice the frequency of fillers of BNC-DEM — there must be additional factors that account for this astonishing difference. BNC-CG is a very heterogeneous collection of texts, and the label "formal" does little to explain the use of fillers. We may venture a guess that speakers in many of the subtexts use longer turns or more complex sentence structures that require greater effort in planning than everyday small talk, and therefore cause speakers to produce fillers to hold the floor or to signal that more is coming, but only a closer qualitative study of the material will help solve this intriguing problem.

This is also true as regards the difference between the LLC and BNC-DEM. LLC speakers are mostly male and highly educated, but again, there must also be other reasons for the much higher ratio of fillers in the LLC. There may also be a higher incidence of pragmatic markers such as *I mean* and *you know* in BNC-DEM, making up the difference. Finally, the type of transcription used in the two corpora may account for some of the discrepancy: the overall transcription principles of the LLC, which includes intonation and other vocal phenomena, would have entailed greater precision in the transcription of fillers or other marginal words.[18] The current lack of an accessible soundtrack for the BNC makes it impossible to ascertain whether this is true, but close discourse study including the use of pragmatic markers may yield further insights. Ultimately, of course, corpora with available soundtracks need to be used for comparison.

This brings us back to the question of terminology and, even more importantly, to the functions of what I have provisionally called 'fillers' to facilitate reference to earlier work. 'Filler' is a rather negative and uninformative default term that is dependent on the ideal of fluency and that says nothing about the discourse functions of these items. Their role in discourse organization has been pointed out by e.g. Swerts et al. (1996) using Dutch material, and Corley et al. (2007: 658) offer experimental evidence that "words preceded by disfluency [sic] were more easily remembered". Kjellmer (2003: 190) points out that "so-called F[illed] P[ause]s [...] help to organize the utterance for the listener, who will more easily realize its structure and its main point and be able to follow the argument". Kjellmer (2003: 181) also suggests that "they should be looked upon in most cases as task-performing elements, employed to bring about certain effects". The term 'hesitation (marker)' used by some researchers is also negatively charged, and I would therefore suggest, in a more positive vein, that they be referred to as 'planners'.

Planning is usually regarded as "a fundamental property of intelligent behavior", and speakers' planning also gives hearers time to figure out what will come next.[19] As shown by Kjellmer (2003) planners frequently precede long and difficult words. Planners thus facilitate comprehension, as also suggested by Kjellmer, by increasing "projectability" (cf. Auer 2005), and they prepare both speakers and listeners for the introduction of new entities into discourse (cf. Grondelaers et al. 2009). They are essential to the management of discourse and should therefore be studied as the discourse — or pragmatic — markers that they are. One reason for their being neglected in important works like Erman (1987), Holmes (1986) or Schiffrin (1987, 1992) is certainly their lowly status as vocalizations and their lack of recognition as words. Like other discourse markers, they are multifunctional — Kjellmer (2003) lists and discusses several functions like highlighting, turn-holding, turn-yielding, and turn-taking, but all of these functions would be compatible with the prototypical function of planning. More work needs to be done on all aspects of the use of planners.

Notes

[*] I thank Sebastian Hoffmann, Britt Erman, the editors of this special issue and two anonymous reviewers for valuable comments on earlier versions of this paper. I am also indebted to Jack Du Bois and Sean Wallis for giving me helpful information about corpora, to Elizabeth Traugott for discussing various points, to Filippo Svalduz for the use of American data from his M.A. thesis, and to the English linguistics seminar of Stockholm University for a lively and inspiring discussion of my data and results. I alone am responsible for any remaining inadvertencies.

[1] In the SBC transcription, the equals sign denotes lengthening and square brackets overlap.

[2] The reverse is not true. Americans sometimes express wonder when they see *er* or *erm* in texts of British origin and tend to pronounce these items with retroflex [r]. *Um* (but not *uh*) is listed in the online version of the *Oxford English Dictionary*; both forms are used by British psycholinguists like Corley & Stewart (2008), and in the Diachronic Corpus of Present-Day Spoken English (DCPSE), which contains over 400,000 words from the LLC and a similar number from the British part of the International Corpus of English (ICE-GB).

[3] It is also worth pointing out here that other vocalizations are included among "real" words in *The American Heritage Dictionary of the English Language* online: *uh-uh* ("no"), *uh-huh* ("yes"), *uh-oh* ("oh dear") (cf. also Tottie 1989).

[4] A case in point is the violent criticism encountered by Caroline Kennedy when she was seeking to be appointed as a U.S. senator to succeed Hillary Clinton; Kennedy was faulted for her overuse of *you know* as well as, to a lesser extent, *uh* and *um*.

[5] Maclay & Osgood (1959: 23) used authentic corpus material drawn from an academic conference but chose to remove utterances shorter than 80 words. The mean utterance length was 309 words, and the corpus is thus very different from normal conversation.

[6] A caveat is in order here: the comparison is based on the assumption that transcribers of all corpora have used comparable transcription principles.

[7] Data from LLC and SWB are based on Clark & Fox Tree (2002), from LOCNEC on Gilquin (2008), and from SBC on Svalduz (2006). All data from BNC are my own.

[8] BNC was searched using BNCweb, the CQP Edition — XML version (cf. Hoffmann et al. 2008).

[9] BNC-DEM: chi-square 174.14, 1 d.f.; BNC-CG: chi-square 748.06, 1 d.f. All calculations of significance are of course based on original data, not frequencies per 100K.

[10] BNC-DEM: chi-square 199.82, 1 d.f.; BNC-CG: chi-square 140.17, 1 d.f.

11. Notice that the age groups covered by cohorts span different numbers of years: the youngest group and the second oldest cover time spans of 15 years, whereas the intervening cohorts represent 10-year spans. The oldest cohort is open-ended.

12. BNC-DEM: chi-square 439.78, 1 d.f.; BNC-CG: chi-square 484.23, 1 d.f.

13. Chi-square 72.13, 3 d.f.

14. Chi-square 163.74, 3 d.f.

15. The proportions refer to the whole corpus. The annotated sample of BNC-CG has 77% men and 23% women.

16. Chi-square 954.31, 1 d.f.

17. The frequency of *uh+um* in the 421,362 words of LLC that are part of DCPSE is even higher than that recorded by Clark & Fox Tree (2002) for the 170,000-word subset that they used, with 2,703/100K (see Table i). This could be due to the fact that many of the texts added in DCPSE do not consist of impromptu conversation.

Table i. Fillers in LLC according to gender, based on DCPSE

	Approx. no. words	No. uh+um	n/100Kwords
Male	280,908	8,197	2,918
Female	140,454	2,614	1,861
Total	421,362	10,811	2,703

18. I owe this observation to an anonymous reviewer.

19. The quote comes from *Wikipedia*: http://en.wikipedia.org/wiki/Planning (accessed May 2010).

References

American Heritage Dictionary of the English Language (The). 1992. A. H. Soukhanov (Ed.). 3rd ed. Boston: Houghton Mifflin.

American Heritage Dictionary of the English Language (The). Online version. Available at: http://education.yahoo.com/reference/dictionary/ (accessed May 2010).

Andersen, G. 2001. *Pragmatic Markers and Sociolinguistic Variation: A Relevance- theoretic Approach to the Language of Adolescents*. Amsterdam/Philadelphia: John Benjamins.

Auer, P. 2005. “Projection in interaction and projection in grammar”. *Text*, 25 (1), 7–36.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober M. F. & Brennan, S. E. 2001. “Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender”. *Language and Speech*, 44 (2), 123–147.

Boulton, A. 2006. “To er is human”. In M. Pereiro & H. Daniels (Eds.), *Le désaccord*. Nancy: AMAES, 7–32.

British National Corpus (BNC), XML Edition. 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Christenfeld, N. 1995. "Does it hurt to say UM?" *Journal of Non-Verbal Behavior*, 19 (3), 171–186.

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

Collins Dictionary of the English Language. 1979. P. Hanks (Ed.). London and Glasgow: Collins.

Corley, M. & Stewart, O. W. 2008. "Hesitation disfluencies in spontaneous speech". *Language and Linguistics Compass*, 2 (4), 589–602.

Corley, M., MacGregor, L. J. & Donaldson, D. I. 2007. "It's the way that you, er, say it: Hesitations in speech affect language comprehension". *Cognition*, 105, 658–668.

Du Bois, J. W., Schuetze-Coburn, S., Paolino, D. & Cumming, S. 1992. *Discourse Transcription*. Santa Barbara: The University of California.

Encarta. http://encarta.msn.com/encnet/features/dictionary/dictionaryhome.aspx (accessed May 2010).

Erard, M. 2007. *Um… Slips, Stumbles, and Verbal Blunders, and What They Mean.* New York: Pantheon Books.

Erman, B. 1987. *Pragmatic Expressions in English: A Study of* You know, You see *and* I mean *in Face-to-face Conversation*. Stockholm: Almqvist & Wiksell.

Fox Tree, J. E. 2002. "Interpreting pauses and *um*s at turn exchanges". *Discourse Processes*, 34 (1), 37–55.

Gilquin, G. 2008. "Hesitation markers among EFL learners: Pragmatic deficiency or difference?". In J. Romero-Trillo (Ed.), *Pragmatics and Corpus Linguistics. A Mutualistic Entente*. Berlin/New York: Mouton de Gruyter, 119–149.

Goldman-Eisler, F. 1961. "A comparative study of two hesitation phenomena". *Language and Speech*, 4 (1), 18–26.

Google Dictionary. http://www.google.com/dictionary (accessed May 2010).

Gries, S. Th. 2009. *Quantitative Corpus Linguistics with R. A Practical Introduction*. New York: Routledge.

Grondelaers, S., Speelman, D., Drieghe, D., Denis, B., Brysbaert, M. & Geeraerts, D. 2009. "Introducing a new entity into discourse: Comprehension and production evidence for the status of Dutch *er* 'there' as a higher-level expectancy monitor". Unpublished manuscript.

Hoffmann, S., Evert, S., Smith, N., Lee, D. & Berglund-Prytz, Y. 2008. *Corpus Linguistics with* BNCweb — *A Practical Guide*. Frankfurt am Main: Peter Lang.

Holmes, J. 1986. "Functions of *you know* in women's and men's speech". *Language in Society*, 15 (1), 1–21.

Kjellmer, G. 2003. "Hesitation. In defence of *er* and *erm*". *English Studies*, 84 (2), 170–198.

Maclay, H. & Osgood, C. E. 1959. "Hesitation phenomena in spontaneous English speech". *Word*, 15 (1), 19–44.

Merriam-Webster Online Dictionary (The). http://www.merriam-webster.com/ (accessed May 2010).

O'Connell, D. C. & Kowal, S. 2005. "*Uh* and *um* revisited: Are they interjections for signaling delay?". *Journal of Psycholinguistic Research*, 34 (6), 555–576.

Oreström, B. 1983. *Turn-taking in English Conversation*. Lund: Liber Förlag.

Oxford English Dictionary. Online version. Available at: http://www.oed.com/ (accessed May 2010).

Quinting, G. 1971. *Hesitation Phenomena in Adult Aphasic and Normal Speech.* The Hague: Mouton.

Rayson, P., Leech, G. & Hodges, M. 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 (1), 133–152.

Schiffrin, D. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.

Schiffrin, D. 1992. "Discourse markers". In W. Bright (Ed.), *International Encyclopedia of Linguistics*. New York: Oxford University Press, 361–363.

Shillcock, R., Kirby, S., McDonald, S. & Brew, C. 2001. "Filled pauses and their status in the mental lexicon". Paper presented at *Diss '01 — Disfluency in Spontaneous Speech, Edinburgh, UK, 29–31 August*.

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

Stenström, A.-B. 1990. "Pauses in monologue and dialogue". In J. Svartvik (Ed.), *The London-Lund Corpus of Spoken English. Description and Research*. Lund: Lund University Press, 211–252.

Stenström, A.-B. & Svartvik, J. 1994. "Imparsable speech: Repeats and other nonfluencies in spoken English". In N. Oostdijk & P. de Haan (Eds.), *Corpus-based Research into Language. In Honour of Jan Aarts*. Amsterdam/Atlanta, GA: Rodopi, 241–254.

Svalduz, F. 2006. "Why do we say *erm?* A corpus-based study of hesitation markers". M.A. thesis. University of Zurich.

Svartvik, J. & Quirk, R. (Eds.) 1980. *A Corpus of English Conversation*. Lund: Gleerup.

Svartvik, J., Eeg-Olofsson, M., Forsheden, O., Oreström, B. & Thavenius, C. (Eds.) 1982. *Survey of Spoken English. Report on Research 1975–81. Lund Studies in English 63.* Lund: Gleerup.

Swerts, M., Wichmann, A. & Beun, R.-J. 1996. "Filled pauses as markers of discourse structure". *Proceedings of the International Conference on Speech and Language Processing*, Philadelphia, 1033–1036.

Tagliamonte, S. A. 2006. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.

Tottie, G. 1989. "What does *uh-(h)uh mean?*". In B. Odenstedt & G. Persson (Eds.), *Instead of Flowers: Papers in honour of Mats Rydén*. Stockholm: Almqvist & Wiksell International, 269–281.

Author's address

Gunnel Tottie
The Department of English
The University of Zurich
Plattenstrasse 47
CH 8032 Zürich
Switzerland

gtottie@mac.com