Variation and change in the use of hesitation markers in Germanic languages

Wieling, Martijn; Grieve, Jack; Bouma, Gosse; Fruehwald, Josef; Coleman, John; Liberman, Mark

Published in:
Language Dynamics and Change

Variation and change in the use of hesitation markers in Germanic languages

In this study, we investigate cross-linguistic patterns in the alternation between UM, a hesitation marker consisting of a neutral vowel followed by a final labial nasal, and UH, a hesitation marker consisting of a neutral vowel in an open syllable. Based on a quantitative analysis of a range of spoken and written corpora, we identify clear and consistent patterns of change in the use of these forms in various Germanic languages (English, Dutch, German, Norwegian, Danish, Faroese) and dialects (American English, British English), with the use of UM increasing over time relative to the use of UH. We also find that this pattern of change is generally led by women and more educated speakers and holds when functional differences between UM and UH are controlled. Finally, we propose a series of possible explanations for this surprising change in hesitation marker usage that is currently taking place across Germanic languages.

1. INTRODUCTION

Two basic *hesitation markers* (also referred to as *fillers* or *filled pauses*) are common in modern Germanic languages: the UM form, which consists of a neutral vowel followed by a final labial nasal, and the UH form, which consists of a neutral vowel in an open syllable. For example, in the English language these forms are generally written as *um* and *uh* in American English and as *erm* and *er* in British English. Similarly, in German a distinction is made between in *ähm* or *öhm* and *äh* or *öh*, whereas in Dutch a distinction is made between *ehm* or *uhm* and *eh* or *uh*. Similar forms exist in all other Germanic languages.

Hesitation markers, including UM and UH, have long been studied in linguistics, primarily because their use has been seen as being directly related to the cognitive processes responsible for the production of speech, specifically marking disfluencies (e.g., Maclay and Osgood, 1959; Goldman-Eisler, 1968; Rochester, 1973; Crystal 1982; Levelt, 1983; Levelt & Cutler, 1983; Schachter et al., 1991). For example, Schachter et al. (1991) found that lecturers in the humanities used more hesitation markers than lecturers in the natural sciences when teaching, but not when being interviewed. They argued that this difference is due to the larger number of words from which a lecturer in the humanities must choose compared to a lecturer in the natural sciences, where technical vocabulary is more strictly defined. Because humanities lecturers have to make more decisions during speech production, they therefore tend to use more hesitation markers. In other words, hesitation markers are seen as marking *disfluency* during language production. This general explanation for the use of hesitation markers has been referred to as the *symptom hypothesis* (de Leeuw, 2007).

Although disfluencies during language production would appear to explain many occurrences of hesitation markers in spoken language, other explanations for the use of UM and UH have been identified. For example, in a series of reaction time experiments, Brennan and Schober (2001) found that hesitation markers were beneficial to comprehension, as listeners were faster to select a target object after a filler was used in the stimulus sentence. Similarly, Fox Tree (2001) showed that UH (but not UM) facilitated the speed with which listeners were able to recognize upcoming words. Fraundorf and Watson (2011) show that hesitation markers improve recall whether or not they predict upcoming discourse boundaries and that no such effect results from coughs of equal duration, ruling out a processing time effect. In contrast to the symptom hypothesis, this type of explanation for the usage of hesitation markers has been referred to as the *signal hypothesis* (de Leeuw, 2007). Still other researchers have pointed out that UM and UH can be used to fulfill various discursive functions (e.g. Swerts, 1998; Rendle-Short, 2004; Tottie, 2014). For example, Swerts (1998) showed that hesitation markers can be used as markers of discourse structure, with hesitation

markers occurring more often with stronger discourse breaks than with weaker discourse breaks. Similarly, Tottie (2014) argued that UM and UH can be used as discourse markers, with a similar meaning as the discourse markers *well* and *you know*.

Linguists have also directly compared the usage of UM and UH. For example, as noted above, Fox Tree (2001) found that UH but not UM facilitated word recognition by listeners. Alternatively, Shriberg (1994) reported that UM was more frequently found in a sentence-initial position than UH in American English, a result that Swerts (1998) replicated based on the analysis of Dutch data. Similarly both Swerts (1998) and Clark and Fox Tree (2002) found that UH tends to be used by speakers to mark minor delays, whereas UM tended to be used to mark major delays. Such findings, however, have not been replicated by all researchers. For example, O'Connell and Kowal (2005) argued that there are no functional differences in the usage of UM and UH based on their analysis of six media interviews of Hillary Clinton. Furthermore, based on a review of previous research, Corley and Stewart (2008) concluded that as there is no evidence that speakers have intentional control over the production of UM or UH (see also Finlayson and Corley, 2012). Differences have also been found in the use of UM and UH across Germanic languages. For example, De Leeuw (2007) reported that whereas English and German speakers preferred UM, Dutch speakers generally preferred UH.

The aforementioned studies have all focused on the different functions of UH and UM from a structural perspective; however, various researchers have also analyzed the effect of various social factors on the choice between these two forms. For example, Rayson et al. (1997) showed on the basis of a corpus analysis of the British National Corpus (BNC) that *er* (i.e. UH) was the second-most characteristic word for male speech and the fourth-most characteristic word for the speech of older (35+) speakers, whereas *erm* (i.e. UM) was the ninth-most characteristic word for people from the upper social class, although they did not directly contrast social patterns in the use of UM and UH. Liberman (2005), however, noted clear gender- and age-related patterns in the use of UH versus UM in corpora of transcribed English-language telephone conversations (i.e. the Switchboard, Fisher Part 1 and Fisher Part 2 collections; Godfrey & Holliman, 1993; Cieri et al., 2004; Cieri et al., 2005). He observed that the use of UH was higher for men than for women and for older speakers than for younger people, whereas the use of UM with a higher for women and younger speakers. In other words, the frequency of UM relative to UH (i.e. the UM/UH ratio) was greater for younger people and women. As commonly interpreted in sociolinguistic *apparent time* studies (e.g., see Labov, 1994), this variation in hesitation marker usage by age suggests that there is a change underway in the English language with the use of UM relative to UH increasing over time. Furthermore, it would appear that this change is being led by women, which is a common finding in variationist sociolinguistics (e.g., see Labov, 2001).

More recently, various other corpus-based studies have analyzed the use of hesitation markers in English and have obtained similar results (see Tottie, 2011 for an overview). For example, on the basis of two sub-corpora of the BNC (i.e. BNC-DEM and BNC-CG), Tottie (2011) showed that women, younger people, and people from higher socio-economic classes had a higher UM/UH ratio than men, older people and people from lower socio-economic classes—a result that once again suggests that UM usage is rising over time, led by women and speakers from higher classes. Similarly, Acton (2011) analyzed the UM/UH ratio in American English based on the relatively recent Speed Dating Corpus (SDC; Jurafsky et al., 2009) and the older Switchboard corpus (SBC; Godfrey and Holliman, 1993) and obtained similar results, with women showing a greater UM/UH ratio than men in both corpora. Based on the Switchboard corpus, Acton (2011) also showed that this pattern persisted at the dialect-region level and when the gender of the hearer was taken into account (i.e. same-gender dyads appeared to show a greater UM/UH ratio than different-gender dyads). He also found that younger people had a greater UM/UH ratio than older people and that the UM/UH ratio was

greater for the more recent SWBC than the SDC and therefore suggested that these results might indicate that a linguistic change is in progress. Similarly, Laserna et al. (2014) analyzed transcripts of conversations collected by 263 American participants from five different studies (Mehl & Pennebaker, 2003a, 2003b; Mehl, Gosling & Pennebaker, 2006; Fellows, 2009; Baddeley, Pennebaker & Beevers, 2013), which were collected via electronically activated recorders carried by the participants for two to three days, allowing for truly spontaneous conversations to be obtained. Laserna et al. (2014) did not explicitly contrast the use of UM and UH in their study, but they reported a significant correlation between gender (male: 1, female: 2) of $r = -.15$ ($p < .05$) for UH, and $r = -.09$ ($p > .05$) for UM. Consequently, they concluded that women showed a lower frequency of use for both UH and UM than men (since the correlation coefficients are negative) (see also Bortfeld et al., 2001). However, as the reduction appears to be greater for UH than UM, this result suggests that women in this study are characterized by a greater UM/UH ratio than men. In addition, Laserna et al. (2014) reported a negative correlation between age and UM use ($r = -.21$, $p < .001$), but not between age and UH use ($r = -.01$, $p > .05$). As the use of UM (but not UH) decreases for older people, this implies that the UM/UH ratio also decreases for older people, which once again implies that a change in English hesitation marker usage is currently underway.

Previous research on social variation in the use of UM and UH in British and American English has therefore repeatedly identified the same basic patterns: younger speakers and women use relatively more UM than UH compared to older speakers and men. This type of pattern is commonly identified in sociolinguistic research and is seen as indicative of a linguistic change in progress (Labov, 1994). The first goal of this paper is therefore to assess whether a change in hesitation markers usage is truly underway in the English language based on detailed quantitative analyses of both longitudinal and apparent time data. Furthermore, because other Germanic languages have comparable hesitation markers, the second goal of this paper is to investigate whether similar patterns of variation and change in the use of UM and UH can be found in other Germanic languages, including Dutch, German, Norwegian, Danish and Faroese.

2. DATA: SPOKEN LANGUAGE CORPORA

To compare patterns of linguistic variation and change in the use of the hesitation markers UM and UH in Germanic language, we analyzed a range of spoken language corpora representing the English, Dutch, German, Norwegian, Danish and Faroese languages. For each of these corpora we generated a primary data set by extracting information about the usage of UM and UH in the corpus as well as a range of social information about each speaker.[1]

2.1. ENGLISH

For the English language, we analyzed five spoken corpora, including three corpora of American English, one corpus covering a wide range of British English dialects, and one corpus of Scottish English.

First, we analyzed the *Switchboard Corpus* of American English (SBC; Godfrey & Holliman, 1993), which contains data from approximately 2,400 two-sided telephone conversations collected in 1990. We extracted all 91,001 tokens of UM (i.e. *um*) and UH (i.e. *uh*) from the corpus, which were produced by a total of 520 different speakers. In addition, we

[1] The data, methods and results associated with this analysis are available for download as supplementary materials at the first author's website (http://www....) and at the Mind Research Repository (http://openscience.uni-leipzig.de).

recorded the position and duration of the hesitation marker and the duration of preceding and following pauses, as well as the age and gender of each speaker, and the total number of words that they contributed to the corpus.

Second, we analyzed *the Fisher Corpus* of American English (Part 1 and Part 2) (FC*;* Cieri et al., 2004; Cieri et al., 2005), which contains transcripts of almost 12,000 telephone conversations collected from 2002 to 2003. We extracted all 19,753 tokens of UM (i.e. *um*) and UH (i.e. *uh*) from the corpus, which were produced by a total of 10,313 different speakers. In addition, we obtained the age, gender and amount of education (in years) of each speaker, and the total number of words that they contributed to the corpus.

Third, we analyzed the *Philadelphia Neighborhood Corpus* (PNC; Labov et al., 2013), which contains transcripts of interviews with from 395 speakers from the Philadelphia area conducted from 1973 to 2013. We extracted all 25,514 tokens of UM (i.e. *um*) and UH (i.e. *uh*) from the corpus, which were produced by a total of 395 different speakers. In addition, we recorded the duration of the hesitation marker and whether a pause occurred before of after the hesitation marker, as well as the year of recording, the age, gender and number of years of schooling of each speaker, and the total number of words that they contributed to the corpus.

Fourth, we analyzed the spoken component of the *British National Corpus* (BNC; Coleman et al., 2012), which contains approximately seven million aligned words recorded in 1993. We extracted all 25,498 tokens of UM (i.e *erm*) and UH (i.e. *er*) from the corpus, which were produced by a total of 960 different speakers. In addition, we recorded the duration of the hesitation marker and the duration of the pause following the hesitation marker, as well as the age and gender of each speaker, and the total number of words that they contributed to the corpus.

Fifth, we analyzed the HCRC Map Task Corpus of Scottish English (HCRC Map Task Corpus, 1993), which contains transcribed speech collected from undergraduates at the University of Glasgow in 1990, who were participating in a map task in which the guide had to explain a route that could be seen on a paper map to the follower who only had a map without the route. We extracted all 1,987 tokens of UM (i.e. *ehm, erm, mm, um*) and *uh* (i.e. *eh, er, uh*), which were produced by a total of 64 different speakers (of which 61 subjects were Scottish). In addition, we recorded the position of the hesitation marker in each utterance, as well as the age, gender, and role (i.e. follower or guide) of each speaker, and the total number of words that they contributed to the corpus.

2.2. DUTCH

For the Dutch language, we analyzed the *Corpus Gesproken Nederlands* (version 2.0) (CGN, 2006), which contains spoken transcribed speech from various sources (e.g., spontaneous conversations, interviews, telephone dialogues) recorded from 1998 to 2004. We extracted all 228,619 tokens of UM (i.e. *ehm, uhm*) and UH (i.e. *eh*, *uh*) from the corpus, which were produced by a total of 3,433 different speakers. In addition, we recorded the position and duration of the hesitation marker, the duration of preceding and following pauses, the preceding and following word, and the part-of-speech of the preceding and following word, as well as the age, gender, education level, nationality (Dutch, Belgian), and level of preparedness (i.e. low for spontaneous speech, high for a televised speech) of each speaker, and the total number of words that they contributed to the corpus.

2.3. GERMAN

For the German language, we analyzed the *Forschungs- und Lehrkorpus Gesprochenes Deutsch* (FLGD; Depperman, 2014), which contains about 100 hours of recorded speech collected from 2005 to 2014. We extracted all 16,221 tokens of UM (i.e. *ähm, öhm*) and UH

(i.e. *äh, öh*), which were produced by a total of 238 different speakers. In addition, we recorded the age and gender of each speaker.

2.4. NORWEGIAN

For the Norwegian language, we analyzed the *Nordic Dialect Corpus and Syntax Database* (NDCSD; Johannessen et al., 2009), which contains approximately 2.8 million words from conversations and interviews collected between 1951 and 2012. We extracted all 47,604 tokens of UM (i.e. *em, EM, m, M, m-m, m_m*) and UH (i.e. *e, E, h-e*) that were tagged as hesitation markers from the corpus, which were produced by a total of 554 different speakers. In addition, we recorded the year of recording, the age group (old: aged 50+, young: aged between 18 and 30) and gender of each speaker, and the total number of words that they contributed to the corpus.

2.5. DANISH AND FAROESE

Finally, for the Danish and Faroese languages, we analyzed the *Faroese Danish Corpus Hamburg* (FADAC; Braunmüller, 2011), which contains 440,000 words collected on the Faroe Islands from 2005 to 2009. We extracted al 4,504 tokens of UM (i.e. *ehm, ehhm, eehm, æhm, ææhm, øøhm*, etc.) and UH (i.e. *eh, ehh, eeh, æh, ææh, øøh*, etc.) from the corpus, which were produced by a total of 57 different speakers. In addition, we recorded the language in which the interview was conducted (Danish, Faroese), the age and gender of each speaker, and the total number of words that they contributed to the corpus.

3. DATA: TWITTER CORPORA

In addition to analyzing various spoken language corpora, we also analyzed the use of UM and UH in English and Dutch Tweets—a written language register that is especially informal and shares several features with spontaneous speech. In a similar domain of computer-mediated communication, Tagliamonte & Denis (2008) found that the usage rates of discourse-pragmatic variables were broadly similar between instant message conversations and comparable spoken language corpora.

3.1. ENGLISH

For English Twitter, we analyzed a corpus of 6 billion words of American Tweets collected by Diansheng Guo of the University of South Carolina in 2013, which only contains tweets where the longitude and latitude of the user at the time of posting is known, as it was designed for the analysis of geolinguistic variation. We extracted the 69,075 tokens of UM (i.e. *um*) and UH (i.e. *uh*) from the corpus that were produced by the 25,852 users who produced at least 1,000 total words and whose username contained an unambiguous male or female name (e.g. *John2002* was designated as male, whereas *Kate_1234* was designated as female). This approach to identifying gender is not perfect, as some names will be misclassified, but we assume that the chances of misclassifications are relatively modest. In addition, we recorded the gender of each user and the total number of words that they contributed to the corpus.

In addition to this primary English Twitter data set, we also generated a secondary spatial data set based on the geocoded information from the corpus. We extracted all 773,155 tokens of UM and UH from the complete corpus and geographically aggregated the results by county (or county equivalent) by calculating the percentage of UM relative to UH (i.e. UM/(UM+UH)) for all tokens that occurred in a given county, based on the longitude and latitude associated with the tweet containing that token. In total, this process yielded an UM percentage for 2,725 out of the 3,110 counties in contiguous United States.

3.2. DUTCH

For Dutch Twitter, we analyzed a corpus of 28.9 billion words of Dutch Tweets collected by the Department of Information Science at the University of Groningen between 2011 and 2014. We extracted the 68,089 tokens of UM (i.e. *uhm, um, euhm, ehm*, etc.) and UH (i.e. *uh, uuh, eh, eeh, euh*, etc.) from the corpus that were produced by the 38,651 users who produced at least 1,000 total words and whose username contained an unambiguous male or female name (as described above) and/or a four digit number ranging between 1930 and 2009, which we used to estimate that user's year of birth. This approach to identifying age also is not perfect, as some names will be misclassified, but we assume that the chances of misclassifications are relatively modest.

4. ANALYSIS[2]

Because the dependent variable for each of the primary data sets is binary (i.e. the use of UM versus UH or the number of tokens of UM versus the number of tokens of UH), we assessed the effect of each of our predictor variables (e.g., age, gender, hesitation marker duration) on the use of UM and UH using mixed-effects logistic regression (Agresti, 2007). By using mixed-effects regression we are taking the structural variability associated with speakers into account (see Baayen, 2008). This is important because some speakers may be more likely to use UM (relative to UH) than others (i.e. a random intercept for speaker). Similarly, the effect of each predictor may vary across speakers. For example, for some speakers a longer duration of the pause following a hesitation marker may be more predictive of the usage of UM than for other speakers (i.e. a by-speaker random slope of the duration of a following pause). Since we are using logistic regression, the estimates need to be interpreted with respect to the logit scale (i.e. the logarithm of the odds of observing UM rather than UH). Positive estimates indicate an increased probability of observing UM together with increasing values of the predictor, whereas negative estimates signal the opposite. An estimate of zero indicates that it has no effect on the probability of observing UM.

For all of the primary data sets except one, we obtained the best-fitting model including only significant predictors and supported random intercepts and random slopes. Predictors and random intercepts and slopes were included if they reduced the Akaike Information Criterion (AIC; Akaike, 1974) by at least 2, compared to the model without the random intercept or slope (see also Wieling et al., 2014 for a similar approach). A reduced AIC indicates that the additional complexity of the model is warranted given the increase in goodness of fit. Due to the large number of predictors in the Dutch data set, however, we did not fit the best model but rather fitted a random-intercepts-only model and assessed if the inclusion of individual random slopes affected the significance of the predictors. We only included predictors that remained significant in all cases in the final model. Given the large number of predictors in this model, we also did not assess all possible interactions.

We assessed the goodness of fit of these models (including the random-effects structure) by calculating the index of concordance *C*, which is known as the receiver operating characteristic curve area 'C' (Harrell, 2001). Values of *C* greater than 0.8 indicate a successful classifier, whereas a value of 0.5 indicates the classifier has no predictive power at all. All models had C values close to or over 0.8.

In addition, we subjected the one secondary spatial data set, which was based on the geocoded American Twitter corpus, to a Getis-Ord G_i spatial autocorrelation analysis (Ord

[2] Given that we analyzed nine independent data sets, we decided to provide a simplified summary of the results for all models together in this section, rather than reporting each individual model. The full details for each model can be found in the supplementary materials, which contains all data, all *R* commands used to generate the models, and all results for each individual model, as well as detailed instructions on how to conduct the analysis.

and Getis, 1995), using a reciprocal spatial weights matrix, which generates a *z*-score for each location indicating the degree to which that location is part of cluster of counties where UM is relatively more common, in which case it is assigned a significant positive *z*-score (e.g, ≥ +1.96), or a cluster of counties where UH is relatively more common, in which case it is assigned a significant negative *z*-score (e.g. ≤ -1.96). Additional information about this procedure is provided in Grieve et al. (2011).

5. RESULTS: SPOKEN LANGUAGE

Table 1 presents the effects (including estimations of effect size: the increase in logits of the dependent variable for the categorical predictors, or per 1 standard deviation increase of the numerical predictors) of the speaker-related predictors that were present in at least two data sets (i.e. gender, age, education level, and year of recording) on the use of UM over UH. Table 1 clearly shows that women are more likely than men to use UM as opposed to UH across all data sets. Similarly, Table 1 shows that younger speakers are generally more likely than older speakers to use UM as opposed to UH; only in the case of the relatively small HCRC Corpus, does the effect of age not reach significance ($p = 0.07$). Table 1 also shows that more educated people are more likely to use UM as opposed to UH in the Fisher Corpus and the Dutch corpus, but that the effect of education in the PNC was non-significant. In addition, the effect of education is much smaller than that of age. Finally, Table 1 shows that the use of UM over UH has increased over real-time in the PNC, the Norwegian Corpus, and in the Dutch corpus. Figure 1 visualizes this result for the three data sets. It should be noted, however, that whereas the PNC (1973-2013) and the Norwegian corpus (1951-2012) each span at least 40 years, the Dutch Corpus only spans 13 years and 90% of the data was recorded between 1999 and 2003. When for the PNC year of recording is excluded from the analysis and instead only year of birth and age are taken into account, the most important predictor clearly is year of birth; the effect of increasing age (i.e. older people are still more likely to use UH) is only minimal ($p = .04$).

Significant interactions (e.g., between age and gender) were identified in some models; however, because these interactions did not change the direction of the general effect (e.g., the age effect was negative for both men and women, but less so for men than for women), we did not explicitly include these interactions in the table below (see, however, supplemental materials). Importantly, these effects were found to be significant, while controlling for the effect of other potential important predictors, such as the duration of the pause before and after the hesitation marker (see Table 3). Also note that for the PNC (and for the SBC, but not for the HCRC, nor the BNC), the predictive value of the duration of the pause after the hesitation marker has diminished for people born in more recent years. This suggests, for these datasets, that younger people are using UM more across the board, and are not simply more frequently signaling longer pauses.

Figure 2 presents four graphs for the American English Switchboard data set, which visualize the relationship between age, gender and the use of UM and UH. The first graph (top-left) plots the proportion of UM over UH (i.e. UM/(UM+UH)) by age (divided into four age groups containing roughly the same number of speakers) and gender. The error bars indicate the 95% confidence interval (i.e. 1.96 standard error below and above the mean). This graph shows a clear increase in the proportion of UM over UH across age groups for both men and women with women consistently showing a higher rate of UM usage than men, although all speakers in this corpus generally prefer UH, with only women in the two youngest age groups approaching 50% UM usage. The second graph (top-right) plots the relative frequency of UM and UH taken together (i.e. total hesitation marker frequency relative to all words in the corpus) by age and gender. This graph shows a clear decline in

hesitation marker usage across age groups for both men and women with men consistently using more hesitation markers than women. The third graph (bottom-left) charts the frequency of UM relative to all words in the corpus by age and gender. This graph shows a clear increase in UM use over age groups with women consistently using UM more frequently than men, although this gap appears to be closing in the youngest age group. Finally, the fourth graph (bottom-right) plots the frequency of UH relative to all words in the corpus by age and gender. This graph shows a clear decrease in UH usage across age groups with men using consistently using UH more frequently than women.

Figure 3 presents the same four graphs for the Dutch data set. Overall, the Dutch results are similar to the English results presented in Figure 2. Particularly, the first graph (top-left) also shows a clear increase in the usage of UM over UH across age groups with women showing a higher proportion of UM over UH than men, while the third graph (bottom-left) shows a clear increase in the relative frequency of UM across age groups with women using UM more often than men. In addition, the fourth graph (bottom-right) shows a decrease in the relative frequency of UH across age groups, especially for women. Despite these similarities, differences between the American English Switchboard data and the Dutch data are apparent. Most notably, whereas hesitation markers in English have been showing a clear decrease in frequency across age groups, the second graph (top-right) shows that there is no clear trend in the overall usage of hesitation markers in Dutch, although the distinction between men and women is similar.

The visualizations for the other data sets, which can be found in the supplemental material, all show similar patterns. Most important, all data sets show an increase across age groups in the use of UM over UH and an increase across age groups in the relative frequency of UM, with younger speakers and women using UM more often than men in both cases. In addition, most data sets show a decrease across age groups in the use of UH. There are, however, differences between the nine data sets. Most notably, the relative frequency of hesitation markers across age groups (i.e. the second graph in Figures 2 and 3) varies considerably across the nine data sets.

Despite generally following the same basic trends, there are also considerable differences in the average overall proportion of UM over UH and the relative frequencies of UM and UH across the nine data sets. These results are summarized in Table 2. For example, the average proportion of UM over UH ranges from 27% to 64% for the five English corpora, compared to 50% in the German corpus, 17% in the Danish corpus, 13% in the Norwegian corpus, and 11% in the Dutch corpus. These differences may reflect register variation both within and across the nine data sets.

Finally, Table 3 presents the effects (again including estimations of effect size) of the hesitation marker-related predictors that were present in at least two data sets (i.e. the duration of the hesitation marker, the duration or presence (for the PNC) of a pause before the hesitation marker, the duration or presence (for the PNC) of a pause after the hesitation marker, the presence of the hesitation marker at the start of the utterance, and the presence of the hesitation marker at the end of the utterance) on the use of UM over UH. Table 3 only presents results for the five data sets for which we were able to include information about the duration and position of hesitation markers and pauses.

Table 1. Effects of subject-related predictors on the choice of UM over UH for all data sets

	Gender: Male vs. Female	Age: Old vs. Young	Education: High/More vs. Low/Less	Year of Recording: Increase vs. Decrease
SBC	F (1.03)	Y (0.6_z - 0.7_z)		
FC	F (1.37)	Y (0.39_z)	More (0.11_z)	
PNC	F (1.31)	Y (1.2_z -1.7_z)	(More) (0.03_z)	Increase (0.54_z)
BNC	F (0.45)	Y (0.45_z)		
HCRC	F (2.30)	(Y) (0.35_z)		
German	F (0.43)	Y (0.94_z)		
Norwegian	F (0.23)	Y (0.65)		Increase (0.35_z)
Danish/Faroese	F (0.59)	Y (0.4_z - 0.6_z)		
Dutch	F (0.5 - 0.9)	Y (0.3_z - 0.6_z)	High (0.15_z)	Increase (0.09_z)

Significant ($p < 0.05$) and non-significant (category name put between parentheses) effects are listed; an empty cell indicates the absence of that predictor in that data set. The values between parentheses indicate the effect size (in terms of logits: the increase in probability of observing UM rather than UH) when the category changes to the one indicated or (when a subscripted *z* is shown) when the value of the numerical predictor increases with 1 standard deviation. A range of values indicates the predictor is involved in an interaction (i.e. the effect of age in the SBC varies based on the hesitation marker being phrase final or not, while the effect of gender and age varies per country for the Dutch data set, and the effect of age varies per language in the Danish/Faroese data set, and for gender in the PNC).

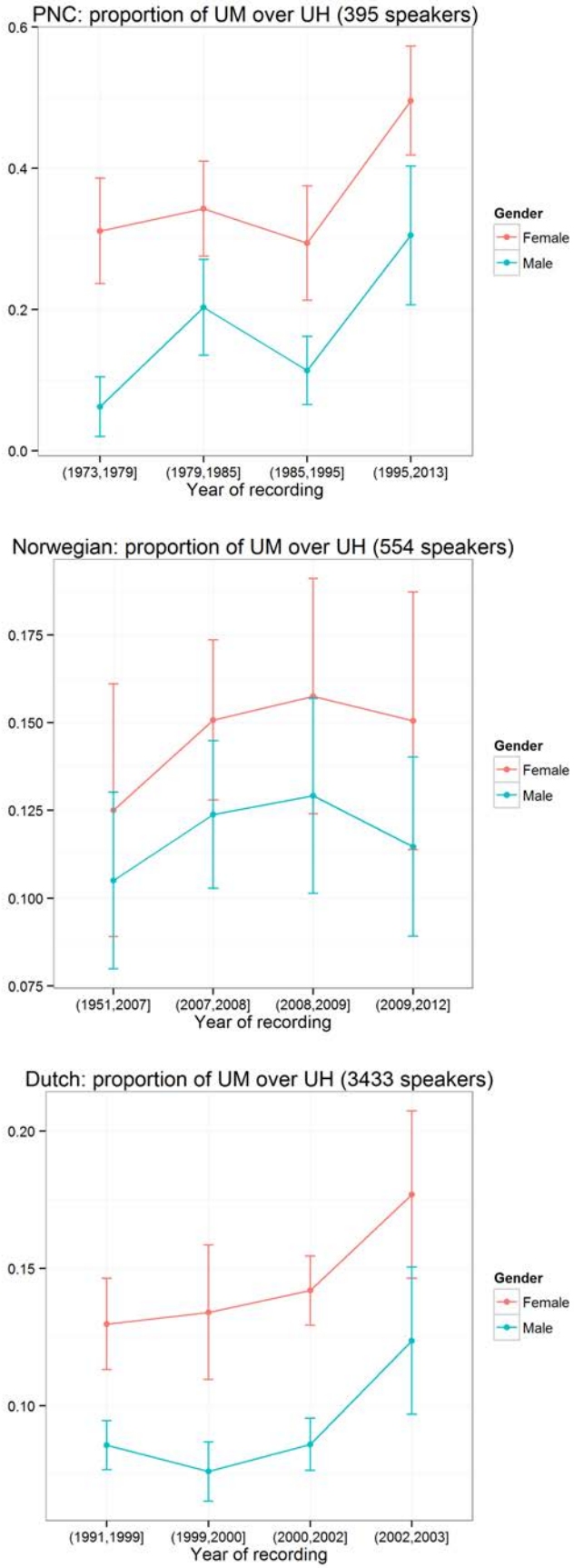

Figure 1. Proportion of UM over UH for three data sets: PNC (top), Norwegian (middle) and Dutch (bottom) by year of recording and gender.

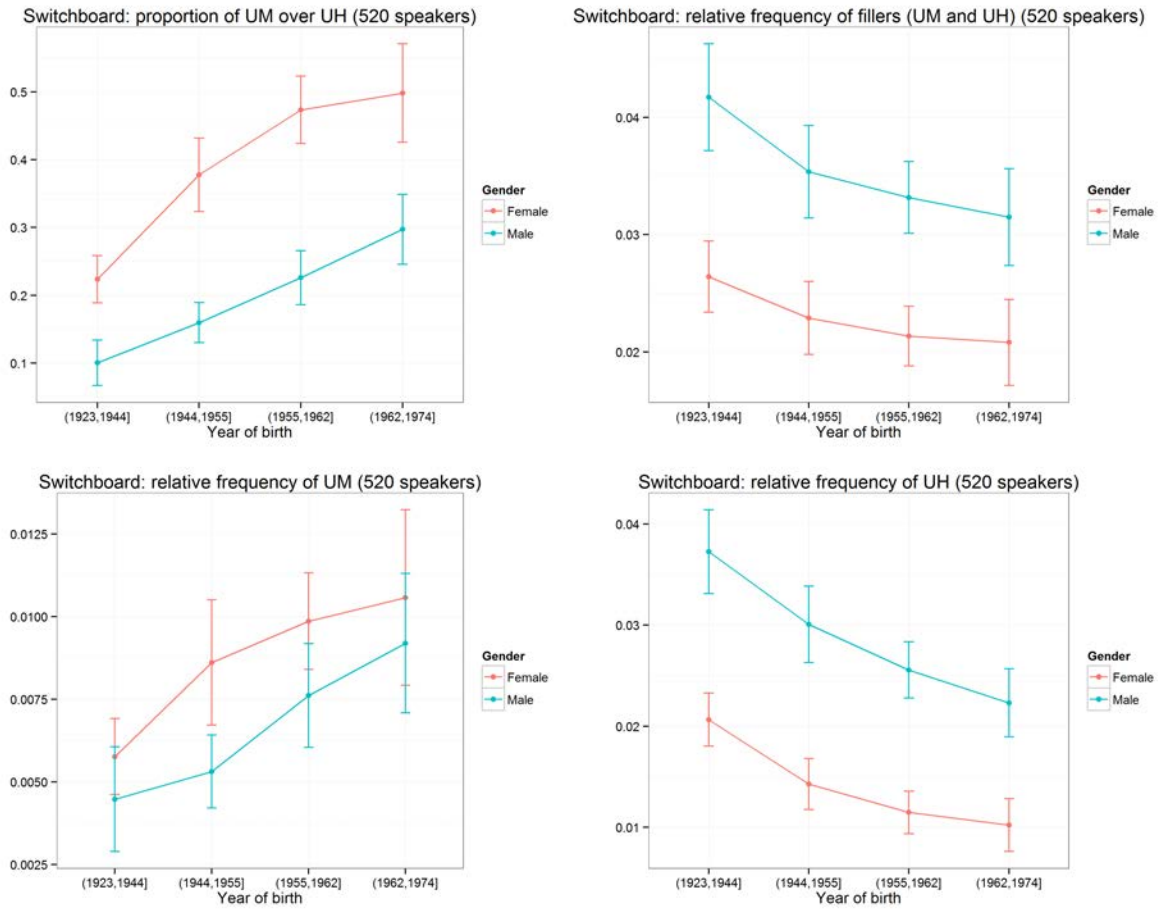

Figure 2. American English Switchboard data: proportion of UM over UH (top-left), relative frequency of hesitation markers (top-right), relative frequency of UM (bottom-left), and relative frequency of UH (bottom-right) by age and gender.

Table 2. Proportion of UM over UH and relative frequency of UM and UH for all data sets

	UM Proportion	UM Relative Frequency	UH Relative Frequency
SBC	0.2825	0.0075	0.0221
FC	0.6408	0.0099	0.0068
PNC	0.2765	0.0045	0.0132
BNC	0.4612	0.0043	0.0045
HCRC	0.5717	0.0081	0.0058
German	0.5017	N/A	N/A
Norwegian	0.1285	0.0026	0.0189
Danish/Faroese	0.1653	0.0020	0.0079
Dutch	0.1086	0.0037	0.0315

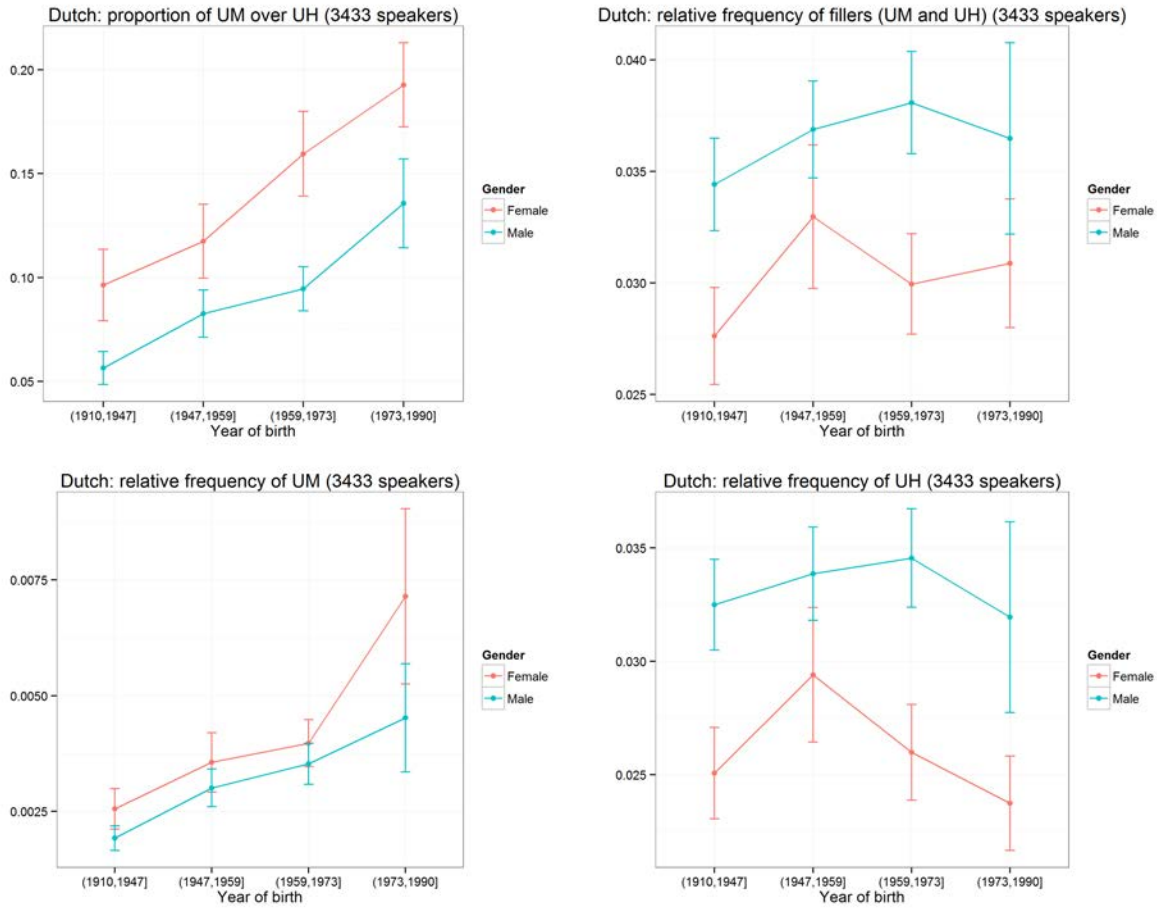

Figure 3. Dutch Spoken data: proportion of UM over UH (top-left), relative frequency of hesitation (top-right), relative frequency of UM (bottom-left), and relative frequency of UH (bottom-right) by age and gender.

Table 3. Effects of hesitation marker-related predictors on the choice of UM over UH

	Duration of Marker	Duration/Presence of pause before Marker	Duration/Presence of pause after Marker	Initial Position	Final Position
SBC	Longer (0.87_z)	Longer (0.12_z)	Longer (0.11_z)	Initial (0.67)	Final (1.06)
PNC	Longer (1.25_z)	(Absent) (-0.08)	Present / Longer (0.59) / (0.55_z)		
BNC	Longer (1.06_z)		Longer (0.44_z)		
HCRC				Initial (0.83)	Final (1.07)
Dutch	Longer (1.15_z)	Longer (0.17_z)	Longer (0.47_z)	Initial (0.51)	Final (0.96)

Significant ($p < 0.05$) and non-significant (category name put between parentheses) effects are listed; an empty cell indicates the absence of that predictor in that data set. The values between parentheses indicate the effect size (in terms of logits: the increase in probability of observing UM rather than UH) when the category changes to the one indicated or (when a subscripted *z* is shown) when the value of the numerical predictor increases with 1 standard deviation.

In general, all predictors showed positive estimates, indicating that higher values of the predictors are associated with a greater likelihood of observing UM as opposed to UH. Specifically, a longer duration (of the hesitation marker or the pause before or after the hesitation marker) is associated with a greater likelihood of the hesitation marker being UM rather than UH, while the occurrence of the hesitation marker in utterance-initial or utterance-final position is also associated with a greater likelihood of the hesitation marker being UM rather than UH. Note that in the case of the Philadelphia Neighborhood Corpus, the presence of a pause before or after the hesitation marker is similar to the hesitation marker being utterance initial or final, as utterances were identified on the basis of the pauses (a pause of 200 ms. or more indicated the break between two utterances).

6. RESULTS: TWITTER

Figure 4 presents four graphs for the American English Twitter data set, which visualize the relationship between gender and the use of UM and UH. The first graph (top-left) plots the proportion of UM over UH and shows that women are more likely to use UM over UH than men. The logistic mixed-effects regression model indicates this effect was significant ($p <$ .001). The second graph (top-right) plots the frequency of UM and UH taken together relative to all words in the corpus and shows that women are more likely to use hesitation markers than men. The third graph (bottom-left) plots the frequency of UM relative to all words in the corpus and shows that women are more likely to use UM overall than men. The fourth graph (bottom-right) plots the frequency of UH relative to all words in the corpus and shows that women are more likely to use UH overall than men. These results for the proportion of UM over UH and the relative frequency of UM agree with the results of the analysis of the American English spoken language data sets (e.g., see Figure 2). However, unlike the results of the spoken analyses, women were found to have higher relative frequencies for UH and for hesitation markers in general, perhaps reflecting functional differences in the use of UM and UH in written language, where true hesitation markers are presumably less common.

Figure 5 presents four graphs for the Dutch Twitter data set, which visualize the relationship between age, gender and the use of UM and UH. The first graph plots the proportion of UM over UH and shows that women and younger tweeters are more likely to use UM than men and older tweeters, although in this case the youngest tweeters were found to reduce their use of UM compared to tweeters from the second youngest group. The logistic mixed-effects regression model indicates that the age effect was significant ($p < .001$) but the gender effect was not ($p = .13$). The second graph plots the frequency of UM and UH taken together relative to all words in the corpus and shows that women and younger tweeters are more likely to use hesitation markers than men, although once again the youngest tweeters were found to reduce their use of hesitation markers compared to tweeters from the second youngest group. The third graph plots the frequency of UM relative to all words in the corpus and shows that women and younger tweeters are more likely to use UM than men, although once again the youngest tweeters were found to reduce their use of UM compared to tweeters from the second youngest group. The fourth graph plots the frequency of UH relative to all words in the corpus and shows that women and younger tweeters are more likely to use UH than men, although once again the youngest tweeters were found to reduce their use of UM compared to tweeters from the second youngest group. In terms of gender, these results all correspond to the results of the analysis of the American Twitter data.

Although the results of the analysis of both the American and Dutch Twitter data correspond well overall with the results of the analysis of the spoken language data sets, the relative frequency of the hesitation markers in the Twitter data is an order of magnitude lower

than in the spoken language data, which likely reflects clear register differences in spoken and written language. Table 4 lists these values, for comparison with the corresponding values for the spoken data sets presented in Table 2.

Finally, in addition to the primary American Twitter data set, a secondary spatial data set was generated consisting of the proportion of UM compared to UH across 2,725 counties in the contiguous United States. Figure 6 maps the raw proportions, where red counties indicate that UM is used relatively more frequently than UH and blue counties indicate that UH is used relatively more frequently than UM. Figure 7 maps the results of the local spatial autocorrelation analysis, which identifies clusters of counties in red where UM is used relatively more frequently and clusters of counties in blue where UH is used relatively more frequently. These maps show a clear regional pattern, identifying the Northeast, Southeast and Upper Midwest as UM regions and the Lower Midwest and most of the West as UH regions.

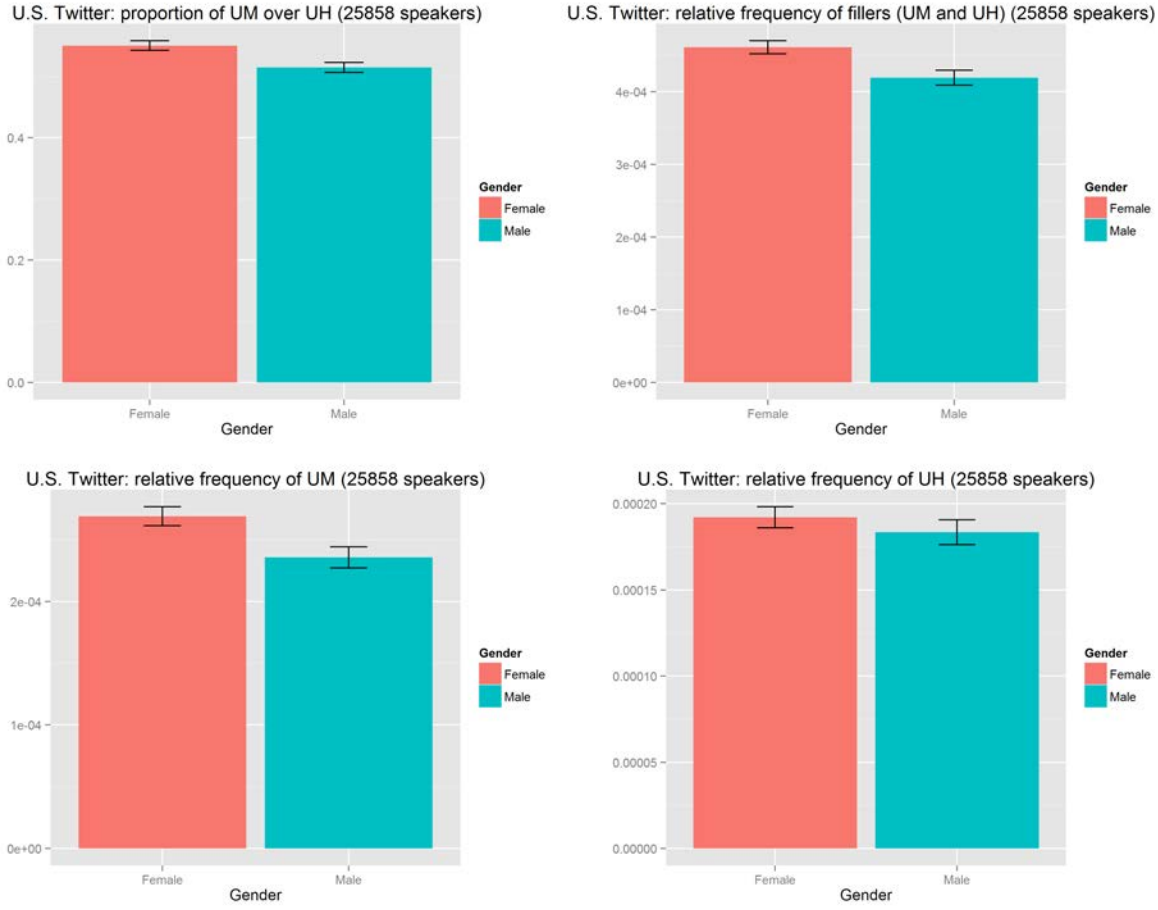

Figure 4. American Twitter data: proportion of UM over UH (top-left), relative frequency of UM and UH (top-right), relative frequency of UM (bottom-left), and relative frequency of UH (bottom-right) by gender.

Table 4. Relative proportion of UM vs. UH and versus all words for the Twitter data sets

	UM Proportion	UM Relative Frequency	UH Relative Frequency
American English	0.5334	0.00025	0.00019
Dutch	0.6518	0.00011	0.00006

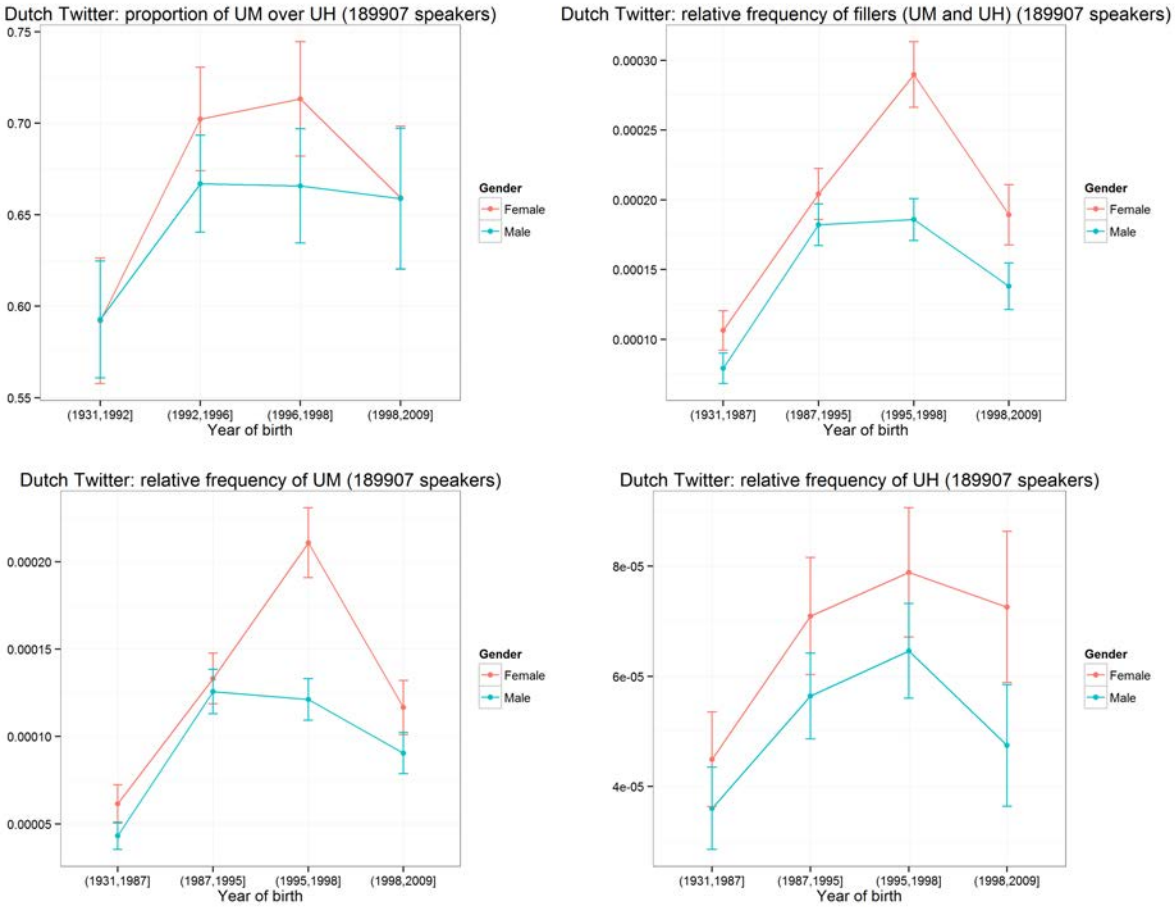

Figure 5. Dutch Twitter data: proportion of UM over. UH (top-left), relative frequency of UM and UH (top-right), relative frequency of UM (bottom-left), and relative frequency of UH (bottom-right) by age and gender.

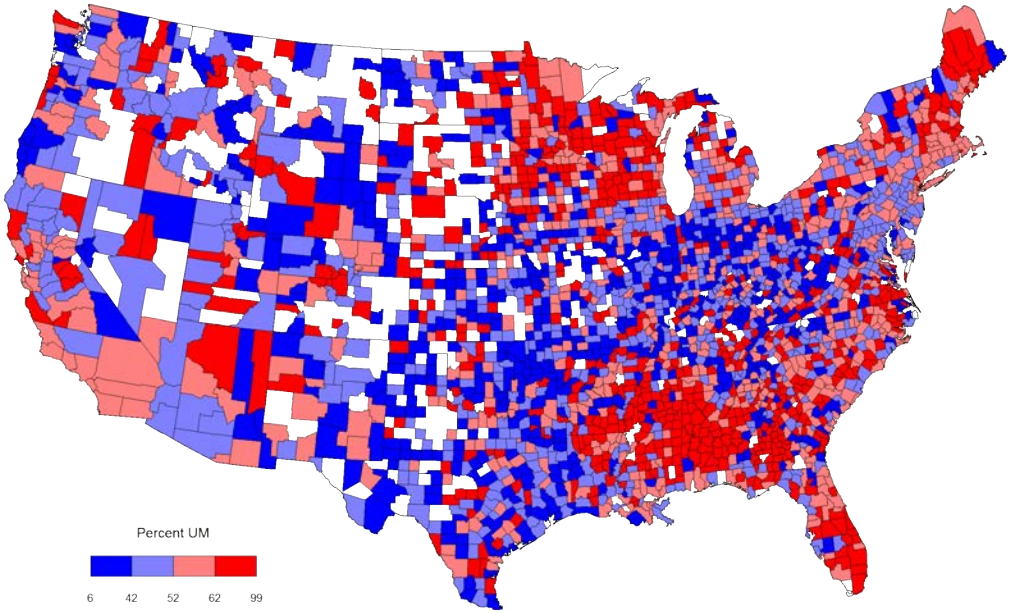

Figure 6. Raw proportions of the use of UM over UH across the United States. Red counties indicate a relatively high proportion of UM clusters, whereas blue counties indicate a relatively high proportion of UH.

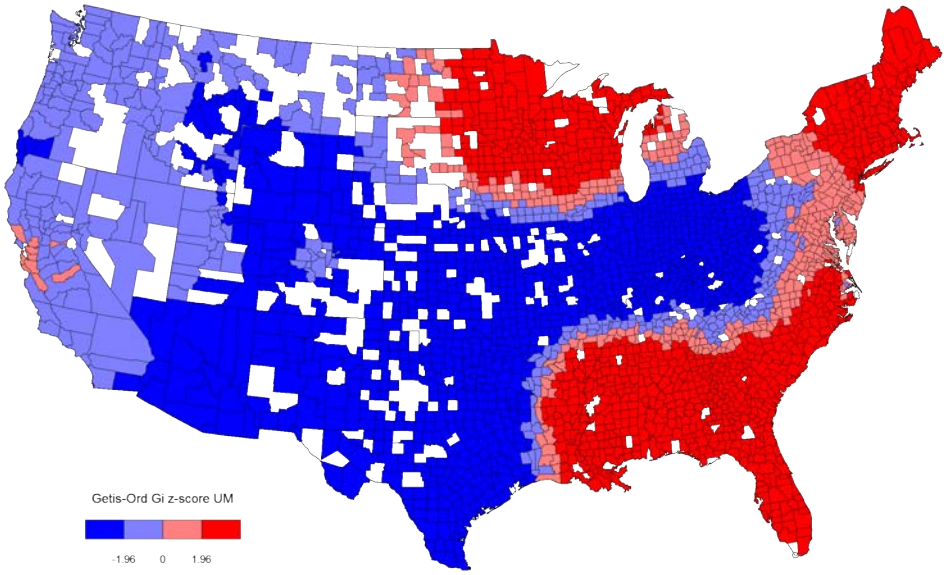

Figure 7. Local spatial autocorrelation analysis of the use of UM over UH across the United States. Red counties indicate UM clusters, whereas blue counties indicate UH clusters.

7. DISCUSSION

The results of our analyses show that there are consistent patterns of sociolinguistic variation in the use of the hesitation markers UM and UH across many modern Germanic languages. In English, Dutch, German, Norwegian, Danish and Faroese, UM is relatively more common in the language of women and younger speakers than UH, which is relatively more common in the language of men and older speakers. Although gender and age patterns in the usage of UM and UH have been identified in previous research on the English language, this paper has shown that this pattern is consistent across a wide variety of Germanic Languages, as well as national dialects of English (American, British, Scottish) and registers of both English and Dutch, including writing.

Given that variation in the use of UM and UH shows a clear trend across age groups, with younger speakers using UM over UH more often than older speakers, it would appear that there is a change in hesitation marker usage currently taking place across various Germanic languages, with the use of UM rising relative to the use of UH over time. Using the age of informants as a way to measure linguistic change is a common technique in sociolinguistic research (see Labov, 1994). This type of *apparent time* analysis is based on the assumption that if a change in progress is taking place, then younger speakers will generally prefer the linguistic form that is on the rise over the linguistic form that is on the decline. Although age-based patterns can also be a result of age-grading, where language gradually changes over the lifetime of individuals regardless of when they were born (e.g., slowing speech rate as people age), this study also found corresponding longitudinal patterns in the Philadelphian English, the Norwegian, and (to a lesser extent) the Dutch data set. The finding that women consistently use UM more often than men across age groups is also consistent with this analysis, as women generally lead linguistic change, especially changes from below (see Labov, 1990) where speakers are not consciously aware of the change, as would appear to be the case for hesitation markers.

This study has therefore uncovered clear evidence that a change is taking place in the use of the hesitation markers UM and UH across a range of mutually unintelligible Germanic languages, with the use of UM becoming more common over time relative to the use of UH. This is a surprising result that cannot easily be explained based on current linguistic theory, especially because this change is occurring simultaneously across so many different languages. In the remainder of this paper, we therefore consider explanations for this cross-linguistic change in the use of hesitation markers.

One possible explanation for this result is that there are independent patterns of change in the usage of UM and UH occurring in all six of the Germanic languages analyzed in this study, with UM *coincidentally* increasing in usage relative to UH in these languages. Such an explanation of cross-linguistic change would seem, however, to be highly improbable. Assuming that this drift is equally likely to progress in either direction (i.e. either toward UM or UH), there is less than a 2% chance that it would progress in the same direction across six languages independently. It is therefore important to consider other hypotheses that directly explain why the same change is taking place across so many mutually unintelligible languages at the same time. In particular, there would appear to be two general types of *non-coincidental* explanations that could account for these results: the change may have spread through contact from one of the languages to the others or a true *parallel change* may be taking place caused by some other factor that affects all of the languages.

Language contact involves linguistic forms spreading across languages through interactions between their speakers. For example, a lexical item in one language that refers to a new concept is often borrowed into other languages that do not have a word to refer to that

concept, such as the English word *computer*, which was borrowed into Dutch, German and Danish, although not into Norwegian (*datamaskin*) or Faroese (*telda*). English forms, in particular, would appear to be especially likely to spread through contact in the modern world due to English being one of the primary languages of mass media and the Internet, as well as being commonly used as a second language, including by speakers of other Germanic languages. Although it is well known that linguistic forms can spread through language contact, it is unclear if language contact could explain the type of cross-linguistic change in hesitation marker usage identified by this study. On the one hand, hesitation markers are relatively common words in the English language, ensuring that they would be present in the language to which non-native English speakers were exposed. The finding that more educated people tend to lead this change is also consistent with spread through language contact, as more educated people are more likely to be second language speakers of English. The usage of UM is also higher on average in the English language corpora compared to the corpora for other Germanic languages, which is what one would expect if the change originated in the English language. On the other hand, there is considerable range in the average usage of UM in the English corpora, which in some cases dips below the levels for German speakers. The use of hesitation markers would also appear to be a highly subconscious process and the shifting usage of UM and UH in the English language is a very subtle change, only having been identified by linguists through the careful statistical analysis of large amounts of language data. Furthermore, unlike the example of language contact presented above, both forms involved in this change already existed in all the Germanic languages under analysis, meaning that it is not the specific form UM that would have been borrowed but a pattern of change that affects a pre-existing alternation. All of these factors presumably make it more difficult for variation in hesitation to spread through contact than, for example, a new word that refers to a new concept. In addition, it is still necessary to explain why the change occurred in English in the first place. Some possible motivations for such a change are discussed below, but a random drift toward UM, as discussed above, does become more plausible if it is assumed that the change began in one language.

In addition to language contact, a change that is taking place across different languages at the same time can also be the result of some general process that causes all the languages to change in parallel. Processes that can lead to parallel changes can be of many types, including phonological change, semantic change, and societal change.

General principles of linguistic change often result in cross-linguistic patterns of change. In particular, many parallel changes are a result of general processes of sound change, such as *elision*, which involves the deletion of segments during speech to facilitate articulation. There does not appear, however, to be any such phonological process that would explain the rise in usage of UM compared to UH over time, such as a tendency for open syllables to close. In fact, the opposite is true: open syllables are generally more common than closed syllables in languages of the world and furthermore syllables consisting solely of a vowel, such as UH, tend to develop onsets as opposed to codas over time (Hyman, 2008). It also seems possible that UM could be reduced to UH through elision in natural speech in order to accelerate language production. It would therefore appear that there are no general phonological reasons to expect that UM would be favored over UH.

Alternatively, a possible semantic explanation for the rise of UM is that there has been a parallel change in the meaning of UM or UH across Germanic languages. To some extent we did control for differences in the use of UM and UH by including various linguistic predictors in our analyses. In particular, we found that UM tended to have a longer duration, to be preceded and followed by longer pauses, and to be more frequently found at the beginning or start of an utterance than UH. These results are in line with earlier studies (e.g., Clark and Fox Tree, 2002, Shriberg, 1994, Swerts, 1998) stating that UM is more likely to signal a major

delay, which is perhaps to be expected given that UM is essentially UH plus the labial nasal; however, the important point is that the gender and age-related pattern still holds when these potential linguistic differences between the two hesitation markers are controlled for. Despite controlling for these linguistic differences, the uses of UM and UH as specific types of discourse markers were not factored into our analyses. As discussed in the introduction, it is clear that hesitation markers can be used as discourse markers—for example to manage turn taking during a conversation or to signal indecision, disagreement or confusion. Consequently, if UM as a discourse marker is becoming more common than UH over time, this could explain the observed trend toward UM in any one of these languages. It would still be necessary, however, to explain why this same change is taking place across six different languages. It is possible that because UM includes a coda it would be a more likely candidate to become lexicalized as a discourse marker cross-linguistically. It is also possible that language contact could be responsible for spreading this change, especially because second language speakers would likely be more aware of a semantic shift in the usage of UM than of a quantitative shift in the alternation between equivalent forms. Nevertheless, as discussed in the introduction, it is unclear if the use of UM and UH as discourse markers differs or is changing over time. In addition, for most of the data sets analyzed here, the overall relative frequency of UM and UH combined was found to be either be decreasing or remaining steady over real and apparent time, which suggests that there has not been a substantial increase in the use of UM or UH as discourse markers over this period of time. Furthermore, most usages of both forms appear to be genuine hesitation markers in the spoken corpora.

Finally, it is possible that this cross-linguistic change in the usage of UM and UH is a result of general societal change. This type of explanation has been used in the past to account for change within languages. For example, Biber et al. (2010) found that noun phrase modification in English newspaper writing has become syntactically more complex and compressed over time and argue that this is due to the increasing amount of information that needs to be incorporated into newspapers in modern times and the increasing use of word processing technology that has allowed reporters to devote more time to carefully preparing and editing their texts. Although their study focused only on English, a similar type of societal explanation might explain the results of this study cross-linguistically. In particular, this change in hesitation marker usage may be due to general cultural and economic changes in the Western World since World War II, including the rise of living standards, education level, and the service economy, all of which may have led to the rise of more formal and polite language production. At the same time, there has been a rise in the use of audio and visual technology, mass communication, and the Internet, all of which may have led to greater attention being paid to language and communication, especially at the personal level, where people now can regularly review their own linguistic production. These large scale societal changes have undoubtedly influenced how we use language, potentially leading to a shift toward more formal, polite, careful, and self-conscious language use in the Western World over the course of the twentieth century. These societal changes could therefore have affected the use of UM and UH in Germanic languages over this time period, as UM is arguably the more formal, polite, careful and self-conscious form than UH. Specially, the pronunciation of UH leaves the mouth in an open position, which is generally considered to be impolite in European society. Furthermore, it would appear that the UH sound as opposed to the UM sound is also a common reaction to physical pain, fatigue, sadness, and anger. Because UM has less physical and negative connotations than UH, it may therefore have become favored over time. Because all of these Germanic languages exist in a similar western reality and have a similar alternation between UM and UH, it is therefore possible that UM is increasing in frequency in all of these language because of a similar general trend toward more refined language use.

This explanation is also consistent with the finding that UM usage is associated with more educated speakers and longer pauses, which may indicate increased care during language production. It is also arguably supported by the finding that UM is more common in the language of women. In addition, this result appears to be supported by the geographical analysis of the American Twitter data. Three main UM cluster were identified in the Northeast, the Southeast, and the Upper Midwest, while the Lower Midwest and the West were identified as UH regions. These results correspond closely to Grieve (2011), where it was found that the Northeast, Southeast, and Upper Midwest are relatively more formal than the rest of the U.S. in terms of contraction rate, with the Northeast and the Southeast using substantially less *not* and verb contraction than the rest of the United States and with the Northeast and Upper Midwest using substantially less non-standard contraction (e.g. *ain't*, *to* contraction, *them* contraction) than the rest of the United States. The regional distribution of UM and UH also appears to be related to historical cultural patterns. In particular, the Northeast and the Southeast were settled early by relatively affluent British colonists, whereas much of the rest of the United States was settled latter by less affluent colonists from Scotland, Ireland and the rest of Europe. The Upper Midwest is also a relatively affluent part of the United States. This result therefore also supports the claim that UM is more formal than UH, as it is preferred in more well-established and more affluent parts of the country.

In conclusion, this study has identified a clear cross-linguistic change in the use of the hesitation makers UM and UH, with UM rising in frequency compared to UH over time across six Germanic languages. A range of possible explanations for this finding were considered, including coincidental change, language contact, phonological change, semantic change, and societal change. Although we cannot fully endorse any of these explanations, two of these hypotheses stood out as being most likely. The first option is that the spread originated in English and spread through contact to other Germanic languages, which naturally have similar forms, possibly reflecting semantic change in the use of UM. The second explanation is that a parallel change is underway due to general societal changes in communication patterns in the Western World, with UM increasing in usage because it is more polite and formal than UH. We do not have sufficient information to choose between these two explanations and nor are we sure that there are no additional explanations that we may have missed, but our results and these hypotheses nevertheless open clear directions for future research. To investigate whether language contact explains these results, it would be necessary to identify other cases where quantitative patterns of change involve pre-existing linguistic forms and to conduct detailed analyses of functional change in the use of UM and UH across Germanic languages. Alternatively, to investigate whether societal change explains these results, it would be necessary to look for general linguistic changes in politeness and formality across Germanic languages in the twentieth century and to investigate whether UM is favored over UH in more polite and formal contexts. Regardless of the results of such analyses, this study as well as any future research it may engender should help extend our understanding of the complexity of language variation change.

REFERENCES

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

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, 19(6), 716-723.

Agresti, A. (200*7*). *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Hoboken, NJ, 2nd edition.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

Baddeley, J. L., Pennebaker, J. W., & Beevers, C. G. (2013). Everyday social behavior during a major depressive episode. *Social Psychological and Personality Science*, 4, 445-452.
Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2), 135-160.
Biber D., Grieve J., & Iberri-Shea H. 2010. Noun phrase modification. In Rohdenburg G. & Schlüter J. (eds.) *One Language, Two Grammars? Differences between British and American English*. Cambridge University Press.
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, 123-147.
Braunmüller, Kurt (2011). Faroese Danish Corpus Hamburg (FADAC Hamburg) [http://corpora.exmaralda.org/sfb_k8.html]
Brennan, S. E., & Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44(2), 274-296.
Cieri, C., Graff, D., Kimball, O., Miller, D., & Walker, K (2004). Fisher English Training Speech Part 1 Transcripts LDC2004T19. Web Download. Philadelphia: Linguistic Data Consortium.
Cieri, C., Graff, D., Kimball, O., Miller, D., & Walker, K. (2005). Fisher English Training Part 2, Transcripts LDC2005T19. Web Download. Philadelphia: Linguistic Data Consortium.
Clark, H. H., & Fox Tree, J.E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84, 73-111.
Coleman, J., Baghai-Ravary, L., Pybus, J., & Grau, S. (2012) Audio BNC: the audio edition of the Spoken British National Corpus. Phonetics Laboratory, University of Oxford. http://www.phon.ox.ac.uk/AudioBNC
Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2(4), 589-602.
CGN (2006). Corpus Gesproken Nederlands. Version 2.0. http://tst-centrale.org/nl/producten/corpora/corpus-gesproken-nederlands/6-17
Crystal, D. (1982). Profiling linguistic disability. London: Arnold.
Deppermann, A., & Schmidt, T. (2014). Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik - Eine exemplarische Untersuchung auf Basis des Korpus FOLK in der Datenbank für Gesprochenes Deutsch (DGD2). In: Domke, Christine & Gansel, Christa (eds.) *Korpora in der Linguistik - Perspektiven und Positionen zu Daten und Datenerhebung* [=Mitteilungen des Deutschen Germanistenverbandes 1/2014], 4-17.
Fellows, M. D. (2009). *An exploration of emotion language use by preschool-aged children and their parents: Naturalistic and lab settings*. PhD thesis. The University of Texas at Austin
Finlayson, I. R., & Corley, M. (2012). Disfluency in dialogue: an intentional signal from the speaker? *Psychonomic bulletin & review,* 19(5), 921-928.
Fox Tree, J. E. 2001. Listener's uses of um and uh in speech comprehension. *Memory and Cognition* 29, 320-326.
Fraundorf, S.H., & Watson, D.G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language* 65(2), 161-175.
Godfrey, J., & Holliman, E. (1993). Switchboard-1 Release 2 LDC97S62. DVD. Philadelphia: Linguistic Data Consortium.
Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. London: Academic Press.
Grieve, J. (2011). A regional analysis of contraction rate in written Standard American English. *International Journal of Corpus Linguistics* 16, 514-546.

Grieve, J., Speelman, D., & Geeraerts, D. (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23, 193-221.
HCRC Map Task Corpus (1993). LDC93S12. Web Download. Philadelphia: Linguistic Data Consortium.
Hyman, L. M. (2008). Universals in phonology. *The linguistic review*, *25*(1-2), 83-137.
Johannessen, J.B., Priestley, J., Hagen, K., Åfarli, T. A., & Vangsnes, Ø. A. (2009). The Nordic Dialect Corpus - an advanced research tool. In: Jokinen, Kristiina and Eckhard Bick (eds.) *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA*. NEALT Proceedings Series Volume 4.
Jurafsky, D., Ranganath, R., & McFarland, D. (2009). Extracting social meaning: Identifying interactional style in spoken conversation. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* ACL, 638–646.
Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*, 2(2), 205-254.
Labov, W. (1994). *Principles of Language Change. Volume 1: Internal Factors.* Oxford: Wiley-Blackwell.
Labov, W. (2001). *Principles of Language Change. Volume 2: Social Factors.* Oxford: Wiley-Blackwell.
Labov, W., Rosenfelder, I., & Fruehwald, J. (2013). One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language*, *89*(1), 30-65.
Laserna, C. M., Seih, Y. T., & Pennebaker, J. W. (2014). Um... who like says you know. Filler word use as a function of age, gender, and personality. *Journal of Language and Social Psychology*, 33(3), 328-338.
de Leeuw, E. (2007). Hesitation markers in English, German, and Dutch. *Journal of Germanic Linguistics*, 19(2), 85-114.
Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, 14, 41-104.
Levelt, W. J. M., & Cutler, A. (1983). Prosodic marking in speech repair. *Journal of Semantics*, 2, 205-217.
Liberman, M. (2005). *Young men talk like old women*. http://itre.cis.upenn.edu/~myl/languagelog/archives/002629.html (first accessed 6 November 2005).
Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, 15, 19-44.
Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90, 862-877.
Mehl, M. R., & Pennebaker, J. W. (2003a). The social dynamics of a cultural upheaval: Social interactions surrounding September 11, 2001. *Psychological Science*, 14, 579-585.
Mehl, M. R., & Pennebaker, J. W. (2003b). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, 84, 857-870.
O'Connell, D. C., & Kowal, S. (2005). Uh and um revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research*, 34, 555-576.
Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, 27, 286-306.
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, 133-152.

Rendle-Short, J. (2004). Showing structure: Using um in the academic seminar. *Pragmatics*, 14(4), 479-498.
Rochester, S. R. (1973). The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research*, 2(1), 51-81.
Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, 60(3), 362-367.
Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*. PhD dissertation, University of California, Berkeley.
Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of pragmatics*, *30*(4), 485-496.
Tagliamonte, S. a., & Denis, D. (2008). Linguistic Ruin? Lol! Instant Messaging and Teen Language. *American Speech*, 83(1), 3–34. doi:10.1215/00031283-2008-001
Tottie, G. (2011). Uh and um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics*, 16(2), 173-197.
Tottie, G. (2014). On the use of uh and um in American English. *Functions of Language*, 21(1), 6-29.
Wieling, M., Montemagni, S., Nerbonne, J., & Baayen, R. H. (2014). Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and socio-demographic variation using generalized additive mixed modeling. *Language*, 90(3), 669-692.