# Before the rise of um

Derek Denis and Timothy Gadanidis

September 9, 2019

## 1 Introduction

One of the most dramatic discourse-pragmatic changes in twentieth-century English has progressed under the radar of laypeople and (until recently) linguists: the rise of *um* as the predominant variant of the 'filled pause' variable (UHM) at the expense of *uh* (Fruehwald, 2016; Tottie, 2011; Wieling et al., 2016). Fruehwald (2016: 43) documents this "textbook" change over 100+ years of apparent time: *um* increases incrementally between generations and the rise is led by women. In this chapter, we investigate (UHM) at an early stage of change to determine what triggered the rise of *um*.

## 2 Um and uh

[əː] [əːm]

## 3 Change in progress

The rise of *um* has now been described extensively in the variationist and corpus-linguistic literature, across a number of corpora and speech communities.

In the British National Corpus, Tottie (2011) observed that *um* was used more frequently than *uh* by women, younger speakers, and more educated speakers; men, older speakers and educated speakers used (UHM) more often overall. Fruehwald (2016)

While these accounts demonstrate definitively that a change is underway, an explanation for the change remains elusive. What was the trigger for this "textbook" change?

In this chapter, we investigate data from before the rise of *um* with the goal of evaluating the functional expansion hypothesis.

## 4 Data

The data for this study are from the *Farm Work and Farm Life Since 1890* oral history collection (Denis, 2016). The corpus consists of oral history interviews with 155 elderly farmers, recorded in 1984. The corpus covers five regions of Ontario, Canada: Temiskaming, Essex, Dufferin, Niagara Region, and Eastern Ontario; for this study, speakers from the latter two regions were considered. Speaker birth years range from 1891 to 1919, just before *um* began to take off per Fruehwald (2016).

We extracted each instance of *uh* and *um* from the transcripts, excluding unrelated instances such as *uh-oh*. Tokens from the two much-younger interviewers was also extracted, and analyzed separately. The transcription protocol emphasized faithful reproduction of *uh* and *um*.

## 5 Coding

We coded for the following social factors: year of birth, gender, and region (Niagara or Eastern Ontario).

To operationalize the functional expansion hypothesis, we coded for utterance position (initial or non-initial).

# 6 Results

## 6.1 Proportional frequency

Table 1 shows how our data compare with previous communities analyzed. The first block summarizes our data from Niagara and Eastern Ontario, as well as F-INT and M-INT, the two younger interviewers. The second block summarizes results from previous work on the Switchboard corpus (Godfrey, Holliman, & McDaniel, 1992), the Fisher corpus (Cieri, Miller, & Walker, 2004), the Philadelphia Neighborhood Corpus (PNC) (Labov & Rosenfelder, 2011), and the British National Corpus (BNC) (2007). The numbers for all of these other corpora are drawn from Wieling et al. (2016).

| Community | Raw N *uh* | Raw N *um* | % *um* | Mean *uh* /1000 | Mean *um*/1000 | Mean UHM/1000 |
|---|---|---|---|---|---|---|
| Niagara | 1864 | 357 | 16.1 | 21.3 | 4.1 | 25.4 |
| E. Ont. | 1563 | 168 | 9.7 | 22.6 | 2.4 | 25.0 |
| F-INT | 321 | 318 | 49.8 | 12.4 | 12.3 | 24.7 |
| M-INT | 255 | 51 | 16.7 | 13.2 | 2.6 | 15.8 |
| Switchboard | — | — | 28.3 | 22.1 | 7.5 | 29.6 |
| Fisher | — | — | 64.1 | 6.8 | 9.9 | 16.7 |
| PNC | — | — | 27.6 | 13.2 | 4.5 | 17.7 |
| BNC | — | — | 46.1 | 4.5 | 4.3 | 8.8 |

Table 1: Cross-community comparison

As can be seen in the table, *um* is less frequent in our data compared to the more recent corpora; the female interviewer uses it around half the time, while the male interviewer's rate is comparable to the farmers'. Relative frequency of (UHM) taken as a whole is on par with other corpora, but we are cautious about making such a comparison because each corpus was collected and transcribed differently (for related discussion, see Pichler, 2010).

Looking at individual speakers' rates, we can see that all speakers use both *uh* and *um*, but there is no clear pattern by age (Figure 1) or gender (Figure 2).

Figure 3 shows the proportion of *um* in apparent time. There is a modest trend upward over time.

Figure 4 shows the pattern when splitting speakers by gender. Starting around 1905, women use *um* slightly more than men do.

Figure 5 shows the pattern when splitting tokens by position (initial vs. non-initial). Starting around 1905, *um* is used more frequently in initial position than in non-initial position.

Figure 6 shows the pattern when splitting tokens by cliticization with *and* or *but* and position.

Figure 7 shows a conditional inference tree for all farmers.

Figure 8 shows a conditional inference tree for the two interviewers.

Taken together, these results show the beginning of the change toward *um* that has been observed by other researchers. While other work has shown that women lead this change, in our data, older women actually use more *um* than the younger women.

Looking at internal factors, we can see that cliticized forms, like *and-uh*, favour *uh*. There is some evidence for positional divergence, possibly consistent with a new utterance-initial discourse function that favours *um* (cf. Fruehwald, 2016, who found no turn-positional difference). Conditional inference trees confirm that the internal constraints persist with the younger speakers, while their baseline *um* rate is higher.

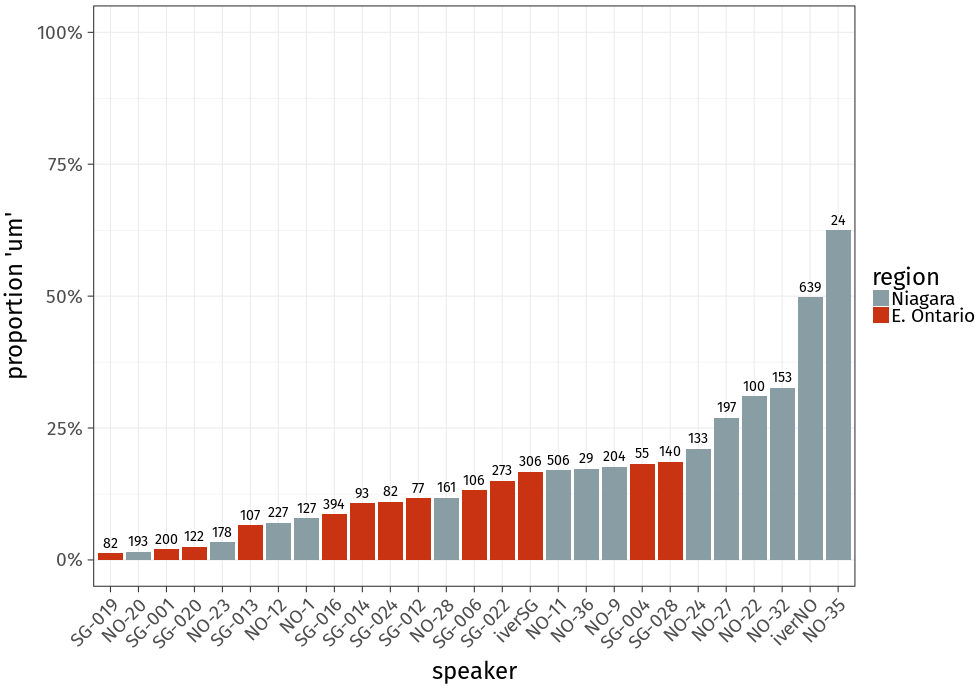


Figure 1: Proportion *um* per speaker by age.

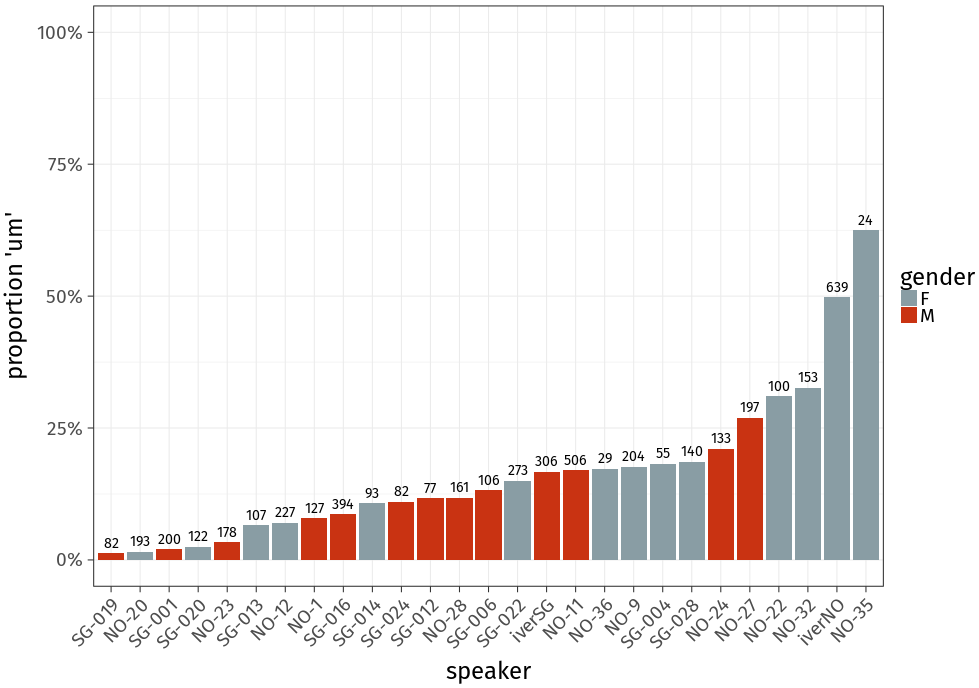


Figure 2: Proportion *um* per speaker by gender.

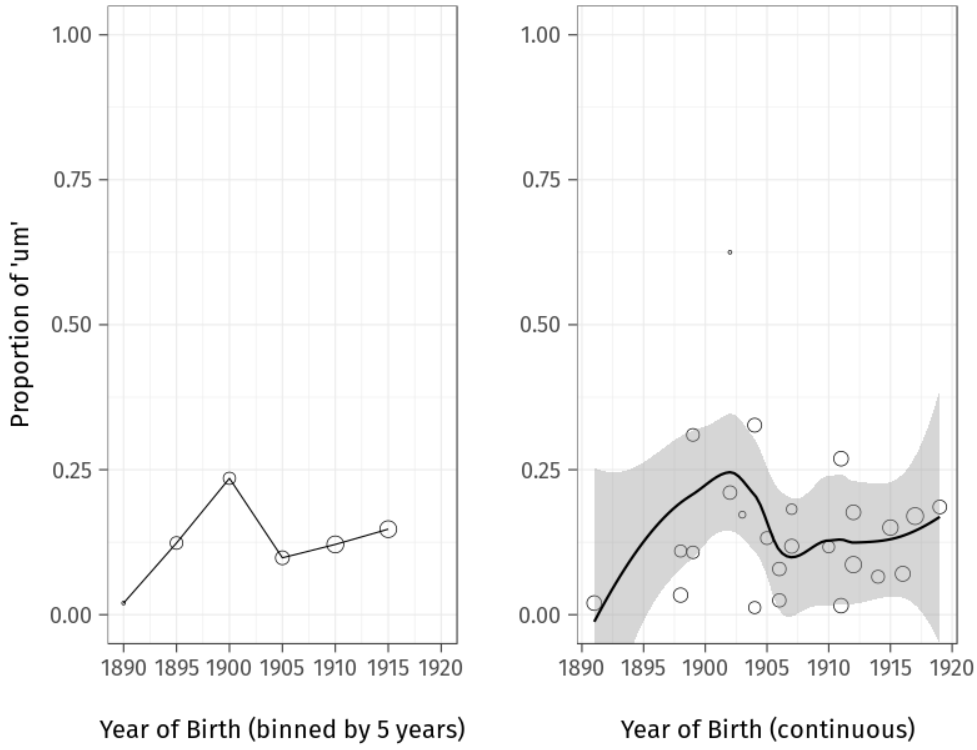


Figure 3: Proportion *um* in apparent time.

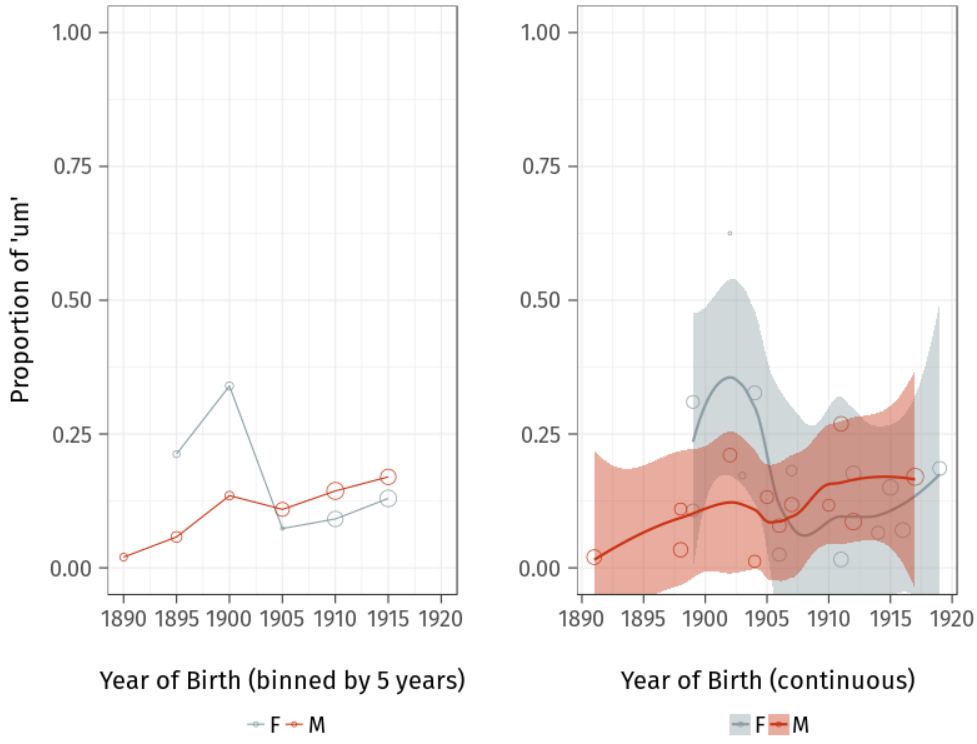


Figure 4: Proportion *um* in apparent time, by gender.

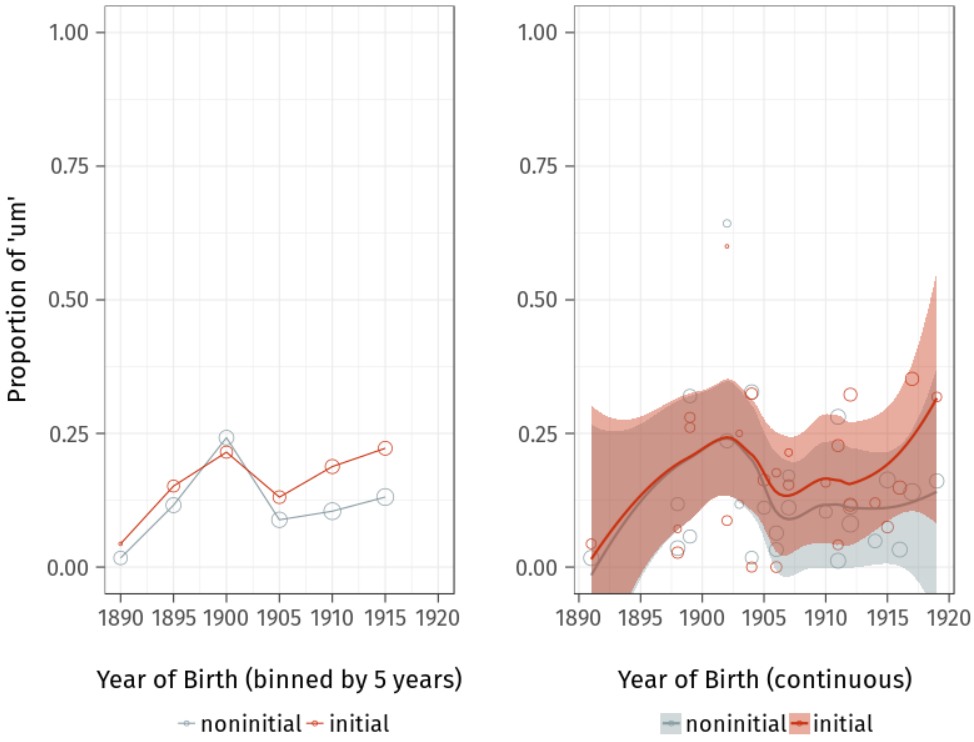


Figure 5: Proportion *um* in apparent time, by position.

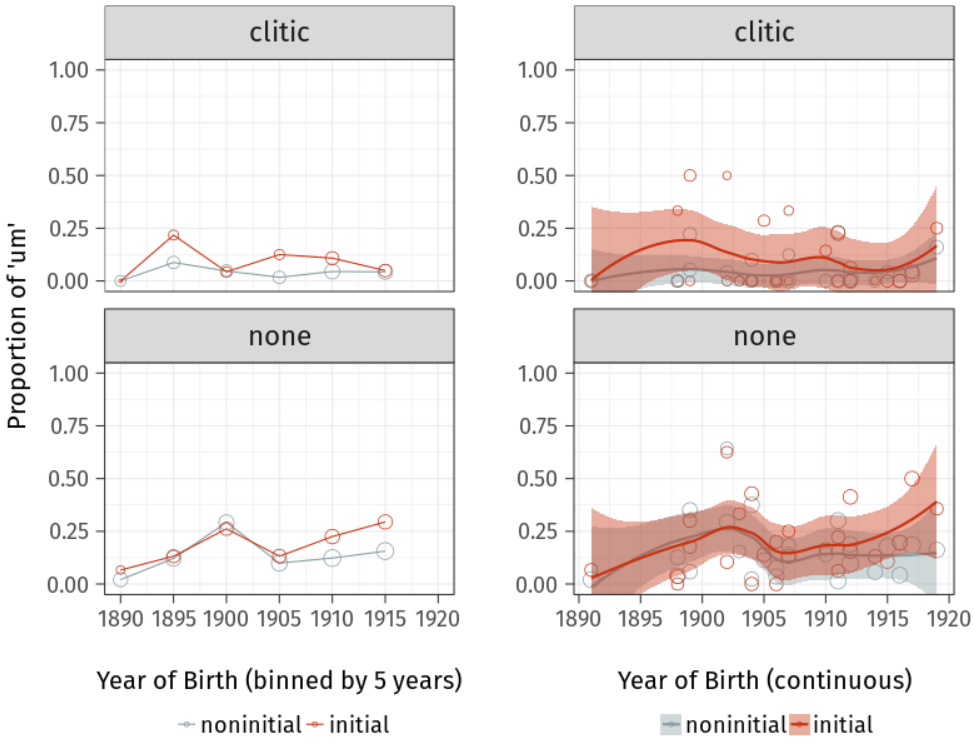


Figure 6: Proportion *um* in apparent time, by position and cliticization.

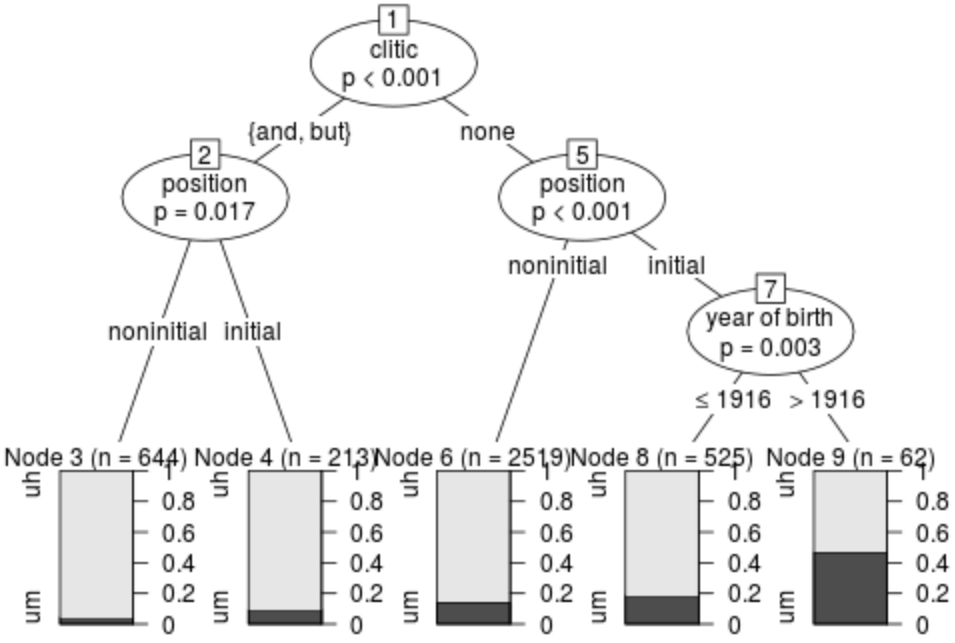


Figure 7: Conditional inference tree for farmers.

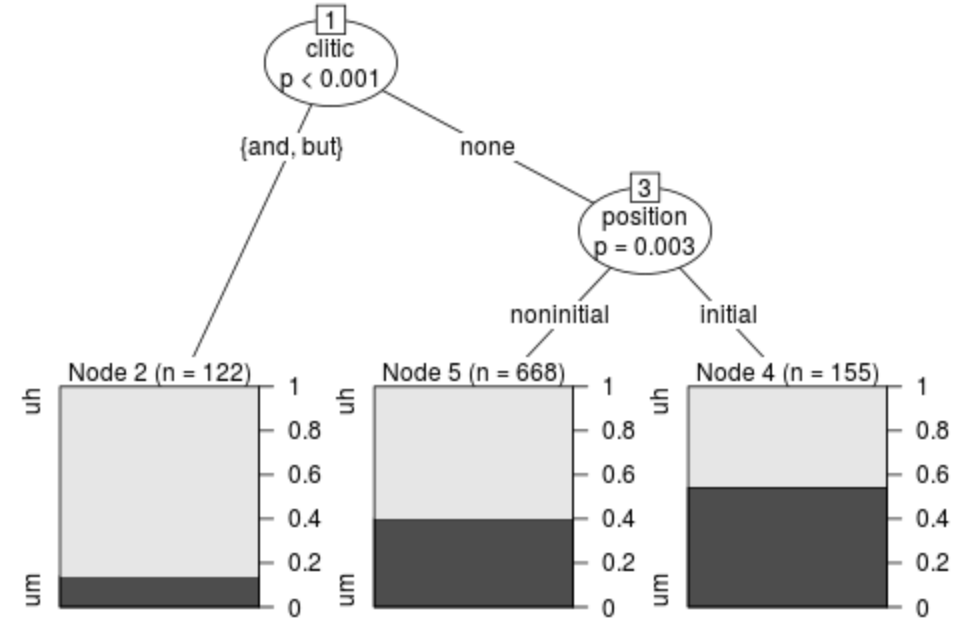


Figure 8: Conditional inference tree for interviewers.

## 6.2 Relative frequency

Fruehwald (2016) tests the hypothesis that functional expansion triggered the rise of um by considering changes to the relative frequency of variants over time (e.g., frequency of *um* or *uh* per 10 000 words). When a new discourse-pragmatic function emerges, we expect that these functions would add to the relative frequency of the feature, and if the new function is restricted to one variant, the relative frequency of that variant should rise, with little change to the relative frequency of the other variant. In other words, we expect a fishtail pattern as with *computer* and *typewriter* over time: once *computer* gained its contemporary meaning, its relative frequency grew additively as that meaning became more frequent. This is illustrated in Figure 9 (Figure 3 from Fruehwald, 2016): looking at the proportion of *computer* over *typewriter* (left graph), *computer* appears to replace *typewriter* over time; but looking at the relative frequency of each word (right graph), it's clear that *typewriter* remained stable as *computer* took off.

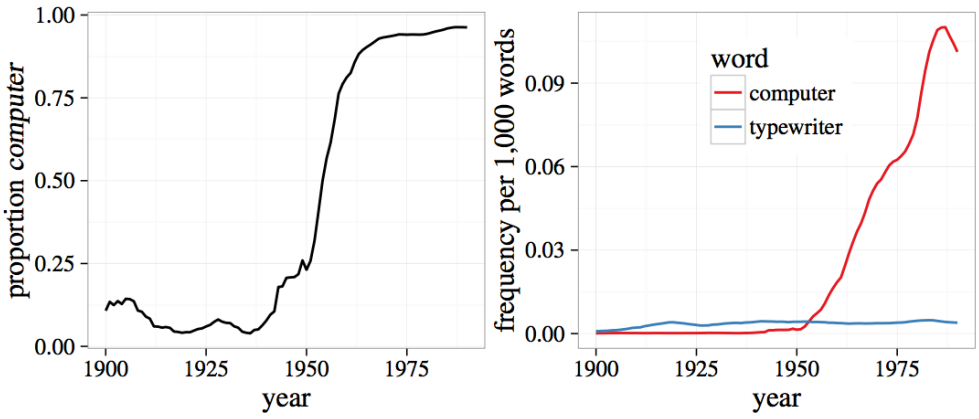


Figure 9: Proportional frequency and relative frequency of *computer* and *typewriter* (Figure 3 from Fruehwald, 2016).

If a new discourse function is what led to the rise of *um*, we should expect to see a similar fishtail pattern, with *um* rising and *uh* remaining stable. Conversely, if *um* were straightforwardly replacing *uh*, we should expect *uh* to fall concurrently with *um*'s rise.

Figure 10 shows the frequency of *um* and *uh* per 1000 words for each of farmers. There is some evidence of a fishtail pattern, but in the opposite direction as expected: *uh* is increasing as *um* remains relatively stable. The pattern is more extreme when we split the data by position, as in Figure 11. In initial position, both *um* and *uh* are largely stable, whereas in noninitial position, *uh* alone is increasing. Splitting the data again by gender, we can see that the increase can be attributed to the female speakers—

# References


Cieri, C., Miller, D., & Walker, K. (2004). The fisher corpus: A resource for the next generations of speech-to-text. In *Lrec* (Vol. 4, pp. 69–71).

Denis, D. (2016). Oral histories as a window to sociolinguistic history and language history: Exploring earlier Ontario English with the Farm Work and Farm Life Since 1890 oral history collection. *American Speech*, *91*(4), 513–516.

Fruehwald, J. (2016). Filled pause choice as a sociolinguistic variable. *University of Pennsylvania Working Papers in Linguistics*, *22*(2), 6.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *[proceedings] icassp-92: 1992 ieee international conference on acoustics, speech, and signal processing* (Vol. 1, pp. 517–520). IEEE.

Labov, W., & Rosenfelder, I. (2011). The Philadelphia Neighborhood Corpus.

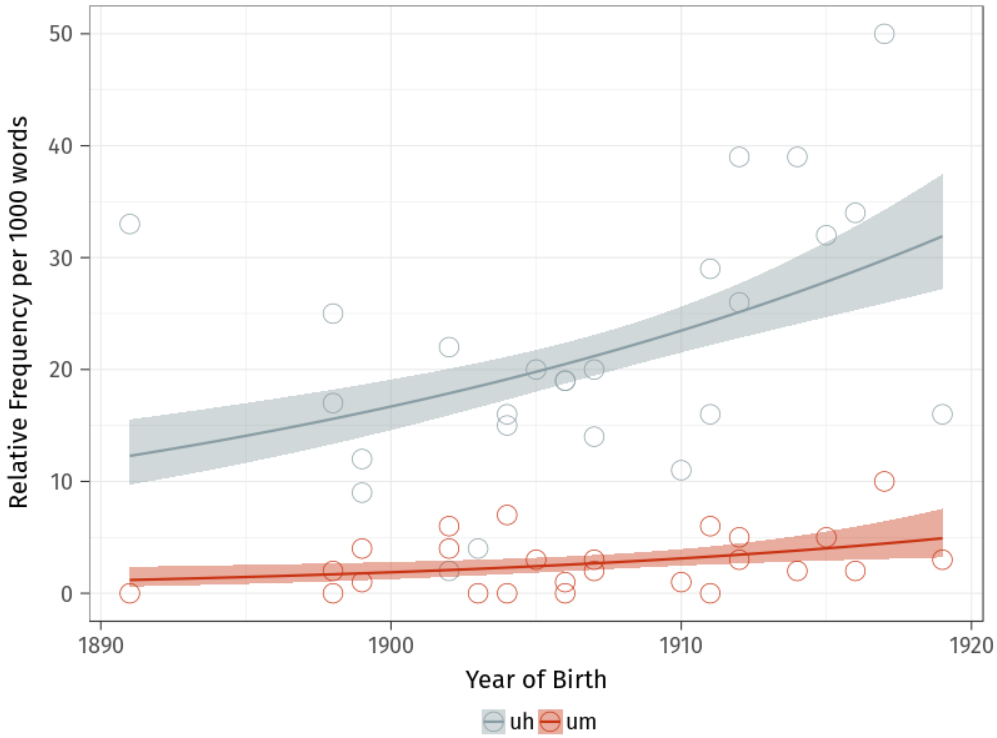


Figure 10: Name

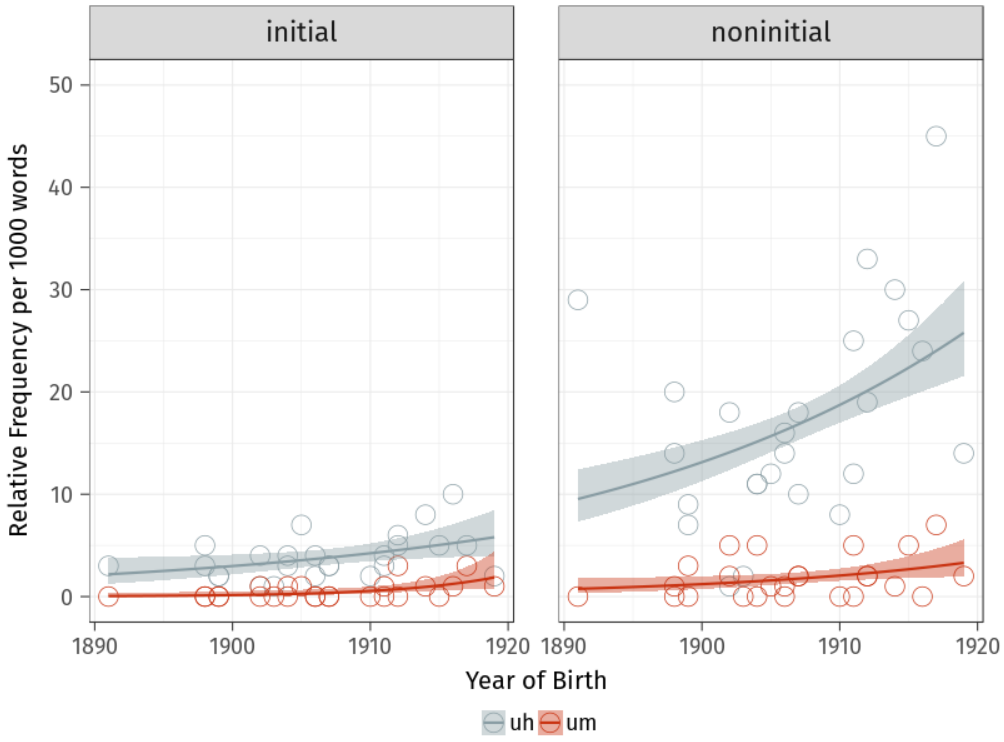


Figure 11: Name

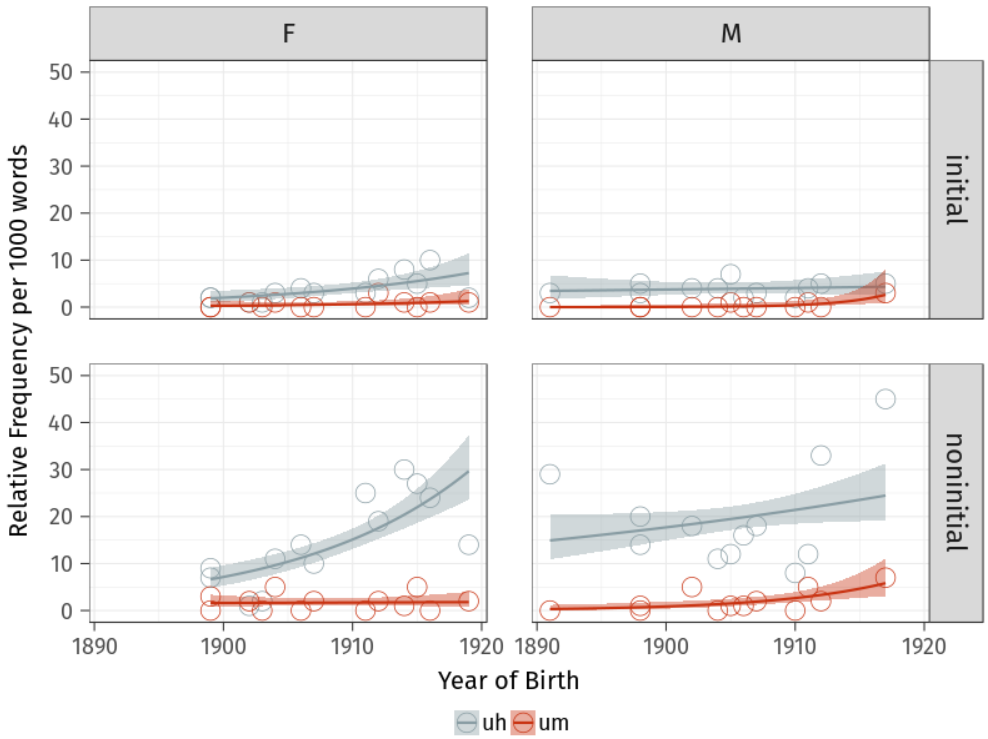


Figure 12: Name

Pichler, H. (2010). Methods in discourse variation analysis: Reflections on the way forward. *Journal of Sociolinguistics*, *14*(5), 581–608.

The British National Corpus. (2007). Version 3. Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium.

Tottie, G. (2011). Uh and um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics*, *16*(2), 173–197.

Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., & Liberman, M. (2016). Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change*, *6*(2), 199–234.