“What’s the ‘uh’ for?”
Pragmatic specialization of UH and UM in IM

Tim Gadanidis

University of Toronto

NWAV47

October 20, 2018

Introduction

The variable

The "filled pauses"/"hesitation markers"/"disfluencies"/… UH and UM, hereafter (UHUM)[1] in instant messaging (IM)

Variants

uh or **um**[2]

(1) a. uh dude, They're having the meeting NOW (M, 1995)

b. um, hostile much? (F, 1986)

[1] /əhʌm/

[2] Also spelled ⟨uhm⟩ by some participants.

(UHUM) as a pragmatic marker

What exactly is (UHUM)?

Views vary; I follow Tottie (2016) who argues that in speech, (UHUM) is a pragmatic marker indicating planning

(UHUM) is used more frequently in word-search, long turns and responses to questions

Change in progress

Both real- and apparent-time data indicate that UM is rising relative to UH (Fruehwald, 2016; Wieling et al., 2016)

Fruehwald (2016), Wieling et al. (2016) suggest that UM may have taken on a new function, leading to its rise, but are unable to identify such a functional difference

(UHUM) in writing

Tottie (2017): in writing, (UHUM) marks stance.

Tottie (2017: 5)

(2) a. **Um, senator**, the market already views those firms as having implicit government backing, because they do … (Paul Krugman, *NYT*, 2010)

b. Obama is more, **um**, seasoned. Barack Obama's … closely shorn hair appears to be increasingly gray. (*Washington Post*, 2010)

Planning as a source for written (UHUM)

Tottie draws a functional difference based on position.

Sentence-initial (UHUM)

"... whereas speakers hesitate to produce answers to questions because they are uncertain of what to say or how to say it, writers merely pretend to hesitate, out of reluctance to say something tactless or hurtful." (Tottie, 2017: 21)

Sentence-medial (UHUM)

"The writer pretends to be searching for a word and pretends to hesitate before making an ironic, funny, somewhat derogatory or naughty choice." (Tottie, 2017: 20)

The present study

Motivation

Tottie (2017) says that (UHUM) is on a lexical cline:

and-uh, *but-uh* clitics in speech on the least wordlike end; stance markers in writing on the most wordlike end

IM is a hybrid register (Tagliamonte, 2016; Tagliamonte & Denis, 2008)—it's conversational and interactive, like speech, but in a written medium

Thus investigating (UHUM) in IM can give us clues to its discourse/pragmatic function and reveal functional differentiation, if it exists

Outline

Outline

- Data and method
- Findings:
    - (UHUM) as a feature
    - (UHUM) as a variable: UH VS. UM
- Discussion
- Conclusion

Data and method

Corpora

TEEN

Data from 11 17–20-year-olds in one social network, 2004–2005, birth years 1985–1987 (Tagliamonte & Denis, 2008)

TEEN

Data from 17 teenagers in Toronto schools, 2004–2006, birth years 1987–1990 (Tagliamonte & Denis, 2008)

FBC

A corpus I built from 9 Toronto-area students in my own community of practice, 2014–2017, birth years 1993–1997

Members of a University of Toronto martial arts club

Predictors

social predictors

year of birth; gender

linguistic predictors

position in message; sentence type (question, response, &c.); polarity; turn-taking

Findings

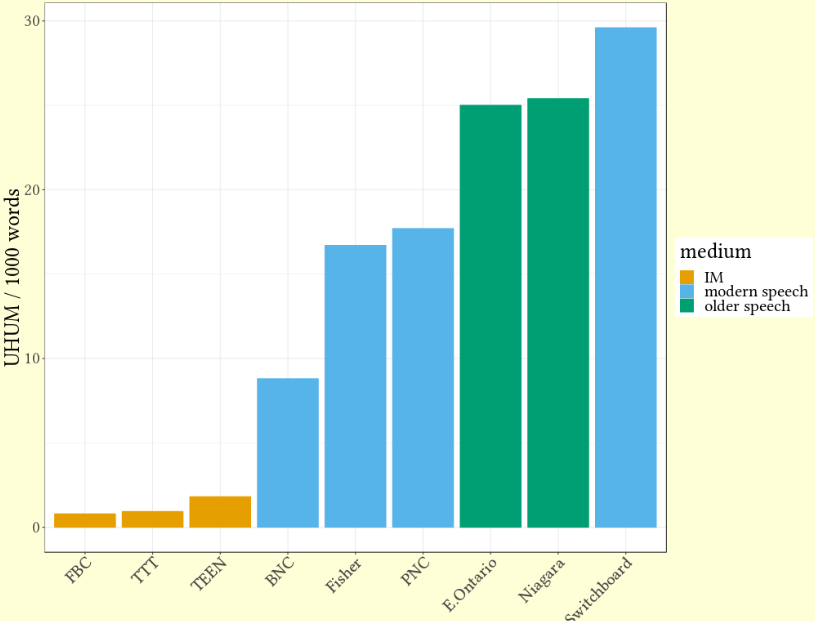
(UHUM) as a feature

Mean relative frequency by corpus

Community	Mean r.f.	Year(s) recorded	Birth years	Source
TTT	.932	2004–2006	1985–1988	current study
TEEN	1.81	2004–2006	1987–1990	current study
FBC	.793	2014–2017	1993–1997	current study
S.board	29.6	1990	1923–1974	Wieling et al. (2016)
Fisher	16.7	2002–2003	???–1986	Wieling et al. (2016)
PNC	17.7	1973–2013	1888–1991	Wieling et al. (2016)
BNC	8.80	1993	???	Wieling et al. (2016)
Niagara	25.4	1984	1898–1917	Denis and Gadanidis (2018)
E. Ontario	25.0	1984	1891–1919	Denis and Gadanidis (2018)

Table 1: Comparison of IM data to historical/contemporary spoken data

Mean relative frequency by corpus



Relative frequency of (UHUM) in IM by year of birth

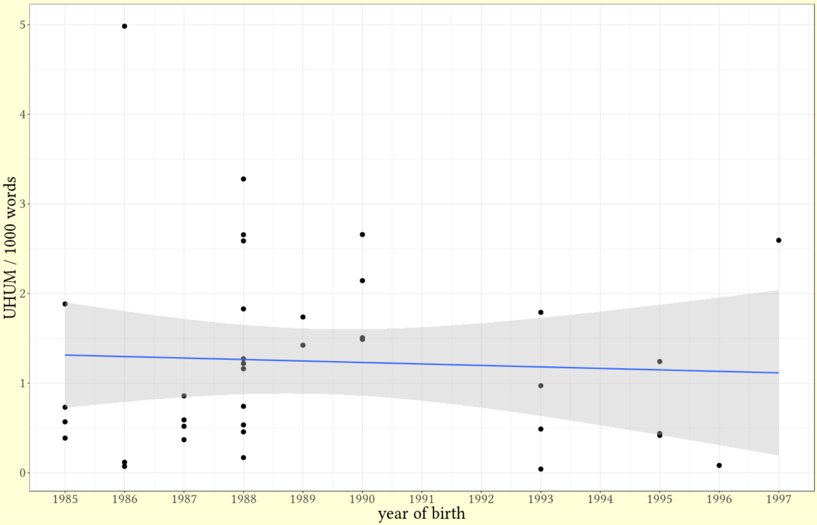


Figure 2: Relative frequency of (UHUM) by year of birth

Predictor variables

Position in message

initial, medial, final, solo (constituting an entire message)

Position in turn

initial, medial, final, solo (constituting an entire turn)

Questions and answers

question, answer, other

Contextual niche

Linguistic contexts in which (UHUM) is most common ($p < 0.01$), based on a generalized linear mixed-effect Poisson regression:

At the beginning of messages and turns[3]

In responses to questions, as opposed to in questions and noninterrogative messages

[3] There's no cumulative effect: an interaction between message and turn position wipes out the effect of turn-initialness when the token is also message-initial.

But what’s it doing?

Overall, (UHUM) appears to be used as a stance marker, whose core meaning is uncertainty (an epistemic stance).

Extract: looking for groceries

1 A: [*sends a picture of the item she wants B to buy*]
2 A: Something like that
3 B: What aisle lol
4 A: **Uhhhh**
5 A: **Uhm**
6 A: Idk
7 A: LOl

UH and UM appear to have somewhat different connotations, as we’ll see in a moment.

Summary

Overall, (UHUM) is a stable discourse-pragmatic feature with an established contextual niche.

Most frequent turn-initially, message-initially and in responses to questions.

No evidence of a change in progress in terms of overall frequency

(UHUM) as a variable: UH vs. UM

Overall distribution

Across all corpora

64% UM; 1513 tokens

Corpus-by-corpus

TTT: 87% UM; 573 tokens

TEEN: 70% UM; 217 tokens

FBC: 45% UM; 723 tokens

Change over time?

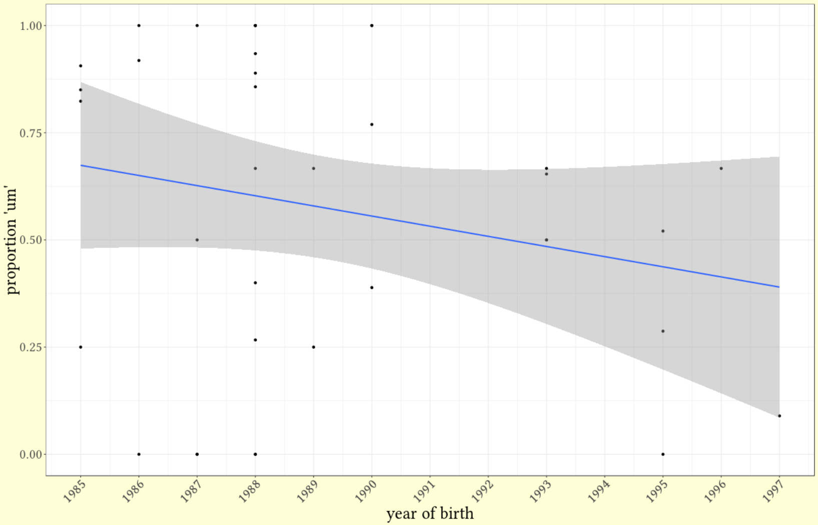


Figure 3: Proportion of UM by year of birth

Individual variation: bars

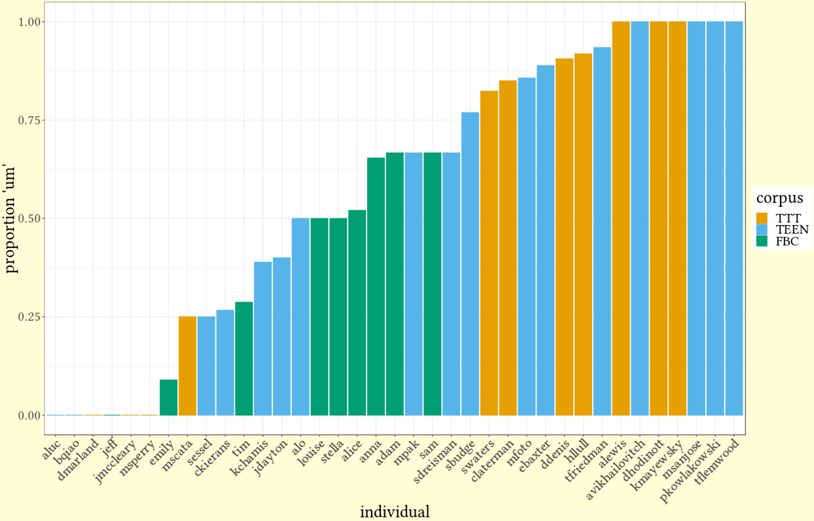


Figure 4: Individuals' rate of UM, sorted

Individual variation: violin plots

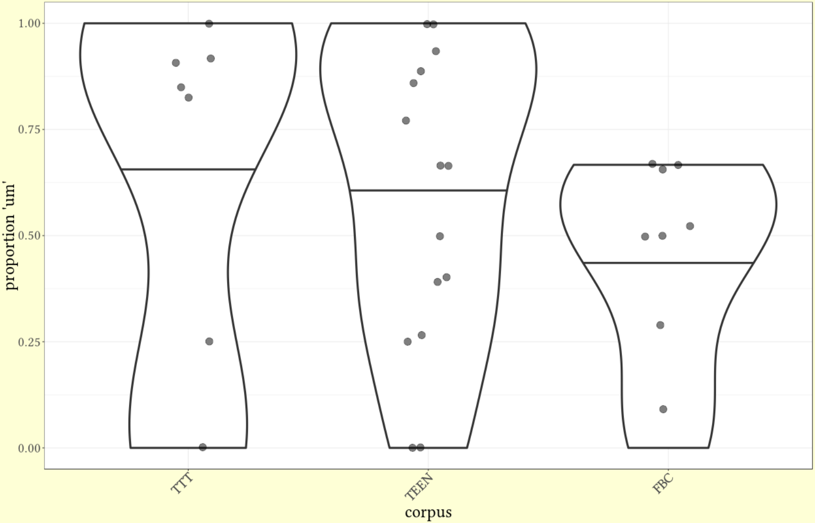


Figure 5: Distribution of individuals' UM rates in each corpus

Variance decreasing

In both TTT and TEEN, we have speakers with 0% um and 100% um

But in FBC, speakers' rates are all between 0% and 66% um and data is more clustered

uh is more frequent and inter-speaker variation is more constrained

Message position

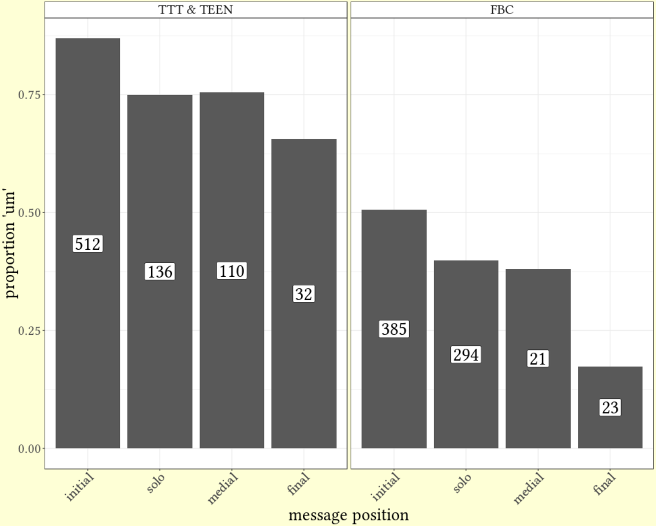


Figure 6: UH VS. UM by message position in each corpus

Questions and answers

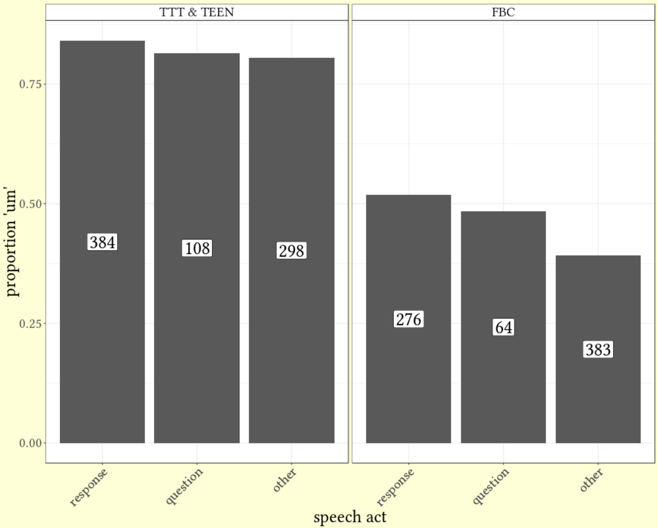


Figure 7: UH VS. UM by sentence type in each corpus

Stronger effects

We see that not only is inter-speaker variance lower in the new data, but the internal constraints are stronger.

It seems we're looking at a developing convention.

A qualitative distinction

UH and UM also appear to have different connotations:

UM has a positively polite connotation and is used to mitigate face challenges/disalignment, while UH is a little more rude (negative politeness) and used to emphasize them.

um: politeness and face-work

Extract: Rice cooker

1 A: **Uhm**, the rice cooker is super hot cuz it was still in keep warm mode o-o

2 B: Holy fuck sorry

3 A: It's okay, let's just be careful next time o.o

uh: challenging and disapproving

Extract: Uh hello

1 A: how did i treat her like a thing
2 B: **uh** hello
3 B: you’ve been trying to change her mind
4 B: trick her into liking you back again

Summary

UH and UM have quantitatively different contextual niches and qualitatively different connotations

Discussion

Reversal of the change in progress?

The IM data is headed the opposite direction from the attested pattern—UH is rising

A possible explanation: specialization (Kroch, 1994)

Specialization

Kroch (1994: 8): competition between members of a doublet will lead to one of two outcomes:

1. one form declines and disappears
2. the forms differentiate in meaning and stabilize

Specialization

Neither variant seems to be disappearing.

So we expect specialization—and that's what we find:

Although they often overlap, the variants are used in different contexts and message positions, and they have qualitatively different functions/indexicalities.

These differences are stronger and there's more constrained variance in the newer data, suggesting the emergence of a convention.

A trajectory

Early state: UH dominant, UM at 5–30% (Denis & Gadanidis, 2018)

UM rises throughout 1900s and early 2000s, reaching up to 64% *um* (Wieling et al., 2016)

Competition between incoming UM and preexisting UH may result in the specialization we see in IM

Some caveats

Register differences

There are crucial differences between speech and IM.

The rate of UHUM in my data is much lower than in spoken corpora, but much higher than in journalism (around 0.0075 per thousand words per Tottie, 2017)

Register differences

In speech, UHUM is used as a planner (Tottie, 2016), but in journalism (Tottie, 2017) and IM, it seems to mark stance.

Due to different communicative needs—spoken utterances are planned in real time, but writing and IM are planned and then sent (in larger or smaller chunks, of course).

All utterances require planning, but not all texts employ overt stance marking, leading to less (UHUM) in writing/IM, and potentially different patterns for each variant.

Register differences and specialization

So how are the IM patterns related to speech?

Two hypotheses, to be investigated in future work:

Continuing the spoken patterns

Specialization we see in IM is a continuation of the spoken patterns, as a result of um's rise and its new competition with uh

Register-limited specialization

Specialization we see in IM is limited to IM, and something else is going on in speech

Either way, I suspect these patterns mostly only apply to the discourse-marker function of (UHUM), which appears to exist in speech (based on my impressions) but is definitely rarer than planning.

Community of practice

How do I know these patterns aren't just exclusive to the FBC community of practice?

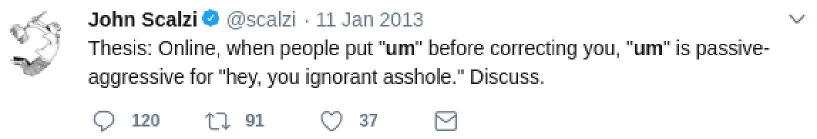
Short answer: I don't.

A community of practice has the ideal conditions for developing this kind of convention: mutual engagement and mutual understanding, taking place over a long timescale.

Community of practice

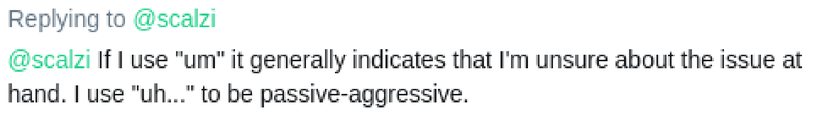
But I think it’s not unlikely that this could be more general.

There’s some metalinguistic evidence, for one thing:



John Scalzi @scalzi · 11 Jan 2013

Thesis: Online, when people put "**um**" before correcting you, "**um**" is passive-aggressive for "hey, you ignorant asshole." Discuss.



Replying to @scalzi

@scalzi If I use "um" it generally indicates that I'm unsure about the issue at hand. I use "uh..." to be passive-aggressive.

Community of practice

Computer-mediated communication also isn't what it used to be— in earlier data, social media (Facebook, Twitter, &c.) was barely a thing and conversations were through one-on-one media like MSN Messenger.

Today though, people's Internet speech practices are more public, so the potential for diffusion could be much higher.

My data don't speak to this issue either way, so analysis beyond speculation will have to wait for future work.

Wrapping up

Summary and takeaways

It looks like UH and UM are specializing in IM: different quantitative patterns and qualitative connotations.

This process seems to be still underway: inter-speaker variation seems to be dropping and internal predictors are strengthening.

(UHUM) is just one part of a developing register of online English which reanalyzes apparently sublexical markers (UM/UH, *hmm*, &c. for stance marking).

Tracking (UHUM) from 2004 to 2017 illustrates the development of a convention for its use as it moves from the spoken domain to the written one.

It remains to be seen whether the patterns I identify here apply in speech as well, and to what extent they exist beyond this community of practice.

Next steps and future research

Expanding the sample to speakers beyond the community of practice

Comparison to spoken data from the same informants (to be collected)

Matched-guise test to test social perceptions of UH vs. UM

Further investigation of apparently nonlexical discourse/pragmatic markers in IM, e.g. *hmm*, *ugh*, where I suspect similar things are going on.

References i

Denis, D. & Gadanidis, T. (2018). Before the rise of *um*. Paper presented at DiPVaC4, Helsinki, Finland.

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

Kroch, A. (1994). Morphosyntactic variation. In *Proceedings of the Thirtieth Annual Meeting of the Chicago Linguistics Society* (Vol. 2, pp. 180–201).

Tagliamonte, S. A. (2016). So sick or so cool? The language of youth on the internet. *Language in Society*, *45*(1), 1–32.

Tagliamonte, S. A. & Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American Speech*, *83*(1), 3–34.

References ii

Tottie, G. (2016). Planning what to say: uh and um among the pragmatic markers. In G. Kaltenböck, E. Keizer, & A. Lohmann (Eds.), *Outside the clause: Form and function of extra-clausal constituents* (pp. 97–122). John Benjamins Publishing Company.

Tottie, G. (2017). From pause to word: *uh*, *um* and *er* in written American English. *English Language & Linguistics*, 1–26.

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.

Acknowledgments

Social Sciences and Humanities Research Council (SSHRC) for research grants 2001–present to Sali Tagliamonte, which enabled the collection of the Tagliamonte Internet Archive, data from which was used in this project

Organizers of NWAV47

University of Toronto Linguistics SSHRC Institutional Grant (SIG) and SGS Conference Grant for sponsoring travel

UofT Linguistics 2017–2018 MA Forum, UofT Language Variation & Change research group, the audiences of CVC10 and DiPVaC4, my supervisor Derek Denis, and Emily Blamire, Marisa Brook, Atiqa Hachimi, Lex Konnelly, F.J. Ruth Maddeaux, Katharina Pabst, Lisa Schlegl, Sali Tagliamonte and Pocholo Umbal for support, feedback and comments

All of the participants who generously volunteered their IM data

Social Sciences and Humanities
Research Council of Canada
Canada

UNIVERSITY OF TORONTO
VARIATIONIST SOCIOLINGUISTICS
LABORATORY

Bonus slides

Mixed-effects model (uh vs. um) i

Predictor	Estimate	Std. Error	z-value	Pr(>|z|)		N-uh	N-um
Overall:							
Intercept	-1.09377	0.88238	-1.240	0.21513		538	975
Message position:							
Initial	*reference level*					257	640
Solo	-0.11703	0.15801	-0.741	0.45891		211	219
Noninitial	-1.07904	0.23431	-4.605	4.12e-06	***	70	116
Year of birth:							
One-year increase	-0.23665	0.10318	-2.293	0.02182	*	N/A	
Gender:							
Female	*reference level*					306	668
Male	-0.73256	0.72060	-1.017	0.30935		232	307
Speech act:							
Non-interrogative	*reference level*					291	390
Question	-0.07558	0.23329	-0.324	0.74595		53	119
Response	0.39255	0.15053	2.608	0.00912	**	194	466
Polarity:							
Negative	*reference level*					83	142
Positive	0.23768	0.19559	1.215	0.22430		455	833

Mixed-effects model (uh vs. um) ii

Random intercept	Variance	Std. Deviation
Individual	5.051	2.247

Table 2: Generalized linear mixed-effects regression model of variation between uh and um with individual as random intercept