Supplementary Material for

Gender and Uptalk in Hong Kong English

by

Wilkinson Daniel Wong Gonzales
The Chinese University of Hong Kong
Chan Pui Yu Ivy
University of Oxford
Zhang Xiaohan Harry
The Chinese University of Hong Kong
Ng Chui Yin Judy
The University of Hong Kong
Chung Yan Ching Karina
The Chinese University of Hong Kong

Part of

Elements in Language, Gender and Sexuality

edited by

Helen Sauntson
York St John University

ISBNs: 9781009634045 (HB), 9781009634069 (PB), 9781009634083 (OC)

Information on this title: www.cambridge.org/9781009634045

DOI: 10.1017/9781009634083

Supplementary resources

A. Glossary of statistical terms

Term	Definition
Bayesian Approach	A statistical method that updates the probability of a hypothesis as more evidence or information becomes available.
Boruta Feature Selection Algorithm	A technique for selecting important features in a dataset by comparing their importance with randomly generated features.
Burn-In or Warm-Up Period	The initial phase in Bayesian simulations or MCMC, discarded to ensure that the results do not depend on arbitrary starting points.
Coefficients	Values in a mathematical equation or model that multiply variables, indicating their influence on the outcome.
Convergence	The state when multiple computational methods or repeated calculations begin to yield similar results, indicating stability and reliability.
Crossover Design (Sampling)	A study design where participants receive multiple treatments in a specific sequence, allowing comparison of the treatments' effects.
Dataset Training	The process of feeding data to a statistical or machine learning model so that it can learn patterns and make predictions.
Effective Sample Size	The number of independent-like observations in a dataset, considering any correlation within the data.
Interaction Factors	Variables in a model that jointly affect the outcome, where the effect of one variable depends on the level of another variable.
Intercept	The starting point or baseline value of a model when all input features are zero.
Logistic Regression	A statistical model that estimates the probability of a binary outcome based on one or more predictor variables.
Markov Chain Monte Carlo (MCMC)	A method used in Bayesian analysis to generate samples from the posterior distribution of parameters.
Mean Decrease Accuracy Method	A technique for determining the importance of variables in a model by evaluating the impact of removing or shuffling them on the model's accuracy.
Mixed-Effects Model	A statistical model that incorporates both fixed effects (common to all units) and random effects (specific to groups or subjects).
Modelling	The process of creating a mathematical representation of a real-world situation to predict or analyze behavior.

Posterior Distribution	The distribution of an unknown quantity, derived from its prior distribution and the likelihood of observed data, in Bayesian analysis.
Probability of Direction	A metric in Bayesian analysis that indicates the likelihood that a parameter has a positive or negative effect.
Random Forests Analysis	A machine learning method that builds multiple decision trees and merges them together to get more accurate and stable predictions.
Random Intercepts	Variables in a mixed-effects model that allow individual subjects or groups to have their own baseline values.
Regression Analysis	A set of statistical processes for estimating the relationships among variables, often used to predict the value of a dependent variable from one or more predictors.
$\hat{R}$	A statistic used in Bayesian analysis to assess the convergence of chains by comparing the variance within and between chains.
Sampling Bias	A bias in which certain members of a population are systematically more likely to be selected in a sample than others.
Sentiment Analysis	The use of natural language processing to identify and categorize opinions expressed in a piece of text, especially to determine the writer's attitude.
Shadow Dataset	A hypothetical or simulated dataset used in machine learning to evaluate model performance under different scenarios.
Slope	The rate at which the dependent variable in a model changes as the independent variable changes.
Weakly Informative Priors	Priors that have minimal influence on the posterior distribution in Bayesian analysis, often used when little prior knowledge is available.
Z-score	A measure of how many standard deviations an element is from the mean of its distribution.

B. Quantitative results: Regression

Table 1. *Bayesian model posterior draw estimates for predictors influencing likelihood to use uptalk or High Rising Terminals (HRT), multi-level categorical variables coded using Weighted Helmert coding conventions, reference levels in boldface.*

Predictors	Median	SD	89% CI (HDI)	*pd*	$\hat{R}$	ESS
Fixed effects						
Intercept	-5.91	22.94	-40.94 – 29.68	0.62	1	3827
Type (opinion vs. **personal experience**)	-0.57	0.11	-0.76 – -0.39	1	1	9847
Type (explanations vs. **opinion/personal experience**)	-0.54	0.13	-0.75 – -0.33	1	1	9568
Type (hard fact vs. **non-hard fact**)	-0.25	0.23	-0.63 – 0.11	0.87	1	6993
Gender identity of speaker (**male** vs. female)	0.05	4.91	-7.54 – 8.1	0.5	1	9883
Socioeconomic status	0.5	2.32	-3.24 – 4.15	0.59	1	3097
Age	0.19	0.84	-1.13 – 1.51	0.6	1	2339
Degree of ethnic identity (Chinese)	-0.03	0.76	-1.19 – 1.21	0.51	1	2398
Degree of ethnic identity (Hong Konger)	-0.31	2.57	-4.39 – 3.78	0.55	1	2738
Gender context (mixed-sex vs. **single-sex**)	0.15	0.33	-0.38 – 0.62	0.7	1	3557
Proficiency (English)	0.05	1.16	-1.82 – 1.91	0.52	1	2357
Institution (CUHK vs. **HKU**)	0.35	3.72	-5.51 – 6.34	0.54	1	3206
Familiarity (**strangers** vs. acquaintances)	-0.07	0.71	-1.25 – 0.99	0.54	1	3944
Sentiment (negative-positive)	0.43	0.21	0.09 – 0.77	0.98	1	8567
Time	-0.01	0	-0.02 – -0.01	1	1	7157
Awareness	0.02	4.93	-7.55 – 8.25	0.5	1	9907
Gender of speaker: Type (**fact** vs. explanation)	0.37	0.25	-0.05 – 0.76	0.92	1	7470
Gender of speaker: Type (personal experience vs. fact/explanation)	0.17	0.3	-0.3 – 0.65	0.71	1	7403
Gender of speaker: Type (opinion vs. non-opinion)	0.82	0.49	0.04 – 1.6	0.96	1	6204

Gender of speaker: Socioeconomic status	0.12	3.16	-4.93 – 5.11	0.51	1	3340
Gender of speaker: Age	-0.21	1.16	-2.06 – 1.63	0.57	1	2161
Gender of speaker: Ethnicity (Chinese)	-0.28	1.39	-2.45 – 1.92	0.58	1	2056
Gender of speaker: Ethnicity (Hong Konger)	0.87	3.09	-4.07 – 5.72	0.61	1	2239
Gender of speaker: Gender context	1.33	0.54	0.53 – 2.18	0.99	1	3530
Gender of speaker: Proficiency (English)	-0.04	1.96	-3.17 – 3.08	0.51	1	2422
Gender of speaker: Institution	-0.96	4.33	-7.88 – 5.92	0.59	1	3552
Gender of speaker: Familiarity	0.13	0.99	-1.36 – 1.72	0.55	1	3333
Gender of speaker: Sentiment	0.67	0.45	-0.04 – 1.39	0.93	1	7838
Gender of speaker: Time	0.01	0.01	0 – 0.02	0.8	1	6141
Awareness: Type (fact vs. explanation)	0.1	0.21	-0.23 – 0.44	0.68	1	7256
Awareness: Type (personal experience vs. fact/explanation)	0.45	0.24	0.06 – 0.82	0.97	1	7322
Awareness: Type (opinion vs. non-opinion)	0.3	0.35	-0.26 – 0.86	0.81	1	8663
Awareness: Socioeconomic status	0.32	3.43	-5.17 – 5.82	0.54	1	3128
Awareness: Age	0.11	1.29	-1.93 – 2.19	0.54	1	3005
Awareness: Ethnicity (Chinese)	-0.53	1.04	-2.16 – 1.16	0.69	1	2809
Awareness: Ethnicity (Hong Konger)	0.26	4.27	-6.7 – 6.92	0.52	1	3084
Awareness: Gender of speaker	0.94	4.2	-5.73 – 7.67	0.60	1	4058
Awareness: Gender context	0.32	0.4	-0.26 – 0.99	0.83	1	3949
Awareness: Proficiency (English)	-0.89	1.43	-3.08 – 1.41	0.74	1	2648
Awareness: Institution	-0.42	3.93	-6.63 – 5.86	0.54	1	3642
Awareness: Familiarity	0.13	1.03	-1.51 – 1.79	0.55	1	4654
Awareness: Sentiment	-0.69	0.37	-1.27 – -0.08	0.97	1	8525
Awareness: Time percent	0	0	-0.01 – 0.01	0.67	1	6555
Random effects						

Conversation pair (Intercept, SD)	0.45	0.27	0.09 – 0.88	1	1	2778
Participant (Intercept, SD)	1.03	1.1	0 – 2.63	1	1	3501
Conversation Pair (1)	-0.07	0.44	-0.77 – 0.56	0.6	1	4549
Conversation Pair (2)	0.19	0.49	-0.48 – 1.01	0.71	1	5024
Conversation Pair (3)	-0.17	0.49	-0.99 – 0.48	0.7	1	5323
Conversation Pair (4)	0.03	0.45	-0.6 – 0.75	0.54	1	6398
Conversation Pair (5)	0.04	0.39	-0.53 – 0.67	0.56	1	5214
Conversation Pair (6)	0	0.43	-0.61 – 0.7	0.5	1	4798
Conversation Pair (7)	0.01	0.46	-0.69 – 0.71	0.52	1	6161
Conversation Pair (8)	-0.33	0.49	-1.17 – 0.3	0.82	1	5978
Conversation Pair (9)	-0.23	0.5	-1.09 – 0.39	0.75	1	5058
Conversation Pair (10)	0.12	0.37	-0.43 – 0.72	0.66	1	5584
Conversation Pair (11)	0.37	0.43	-0.18 – 1.08	0.87	1	4445
Conversation Pair (12)	0.01	0.42	-0.65 – 0.65	0.51	1	7671
Conversation Pair (13)	-0.04	0.47	-0.81 – 0.62	0.55	1	5346
Conversation Pair (14)	0.36	0.54	-0.29 – 1.3	0.82	1	5043
Conversation Pair (15)	-0.31	0.45	-1.05 – 0.28	0.84	1	4329
Conversation Pair (16)	0.01	0.45	-0.67 – 0.72	0.52	1	7279
Participant (Cecilia)	-0.01	1.64	-2.46 – 2.26	0.51	1	7111
Participant (Emery)	0	1.6	-2.41 – 2.26	0.51	1	4949
Participant (George)	0.01	1.6	-2.13 – 2.5	0.52	1	7345
Participant (Gabriella)	0	1.6	-2.38 – 2.35	0.5	1	7956
Participant (Gia)	0	1.52	-2.31 – 2.29	0.51	1	6183
Participant (Hayley)	0	1.6	-2.32 – 2.42	0.5	1	7586
Participant (Johnson)	-0.01	1.66	-2.52 – 2.31	0.51	1	9047
Participant (Levi)	-0.01	1.64	-2.53 – 2.29	0.51	1	7502
Participant (Mason)	0	1.61	-2.48 – 2.27	0.5	1	8853
Participant (Maya)	0.02	1.6	-2.13 – 2.6	0.52	1	7254
Participant (Samuel)	0.02	1.57	-2.15 – 2.49	0.53	1	7444
Participant (Tamara)	-0.02	1.61	-2.39 – 2.34	0.52	1	6783
Participant (Theodora)	-0.02	1.66	-2.55 – 2.21	0.52	1	8504
Participant (Tyson)	-0.04	1.39	-2.31 – 1.86	0.54	1	5573
Participant (Yasmin)	-0.01	1.62	-2.41 – 2.31	0.51	1	7179

Table 2. *Distribution of* UPTALK *variants by variable.*

Variable	**Level**	**[-uptalk]**	**[+uptalk]**
Time	early	87.36% (954)	12.64% (138)
	mid	85.81% (937)	14.19% (155)
	late	86.95% (1,899)	13.05% (285)

Participant	Cecilia	88.32% (174)	11.68% (23)
	Emery	96.48% (438)	3.52% (16)
	George	80.91% (195)	19.09% (46)
	Gabriella	76.83% (199)	23.17% (60)
	Gia	88.58% (349)	11.42% (45)
	Hayley	84.19% (181)	15.81% (34)
	Johnson	95.35% (164)	4.65% (8)
	Levi	90.10% (373)	9.90% (41)
	Mason	99.33% (297)	0.67% (2)
	Maya	67.36% (163)	32.64% (79)
	Samuel	72.46% (242)	27.54% (92)
	Tamara	92.02% (196)	7.98% (17)
	Theodora	95.38% (227)	4.62% (11)
	Tyson	89.06% (415)	10.94% (51)
	Yasmin	76.96% (177)	23.04% (53)
Conversation pair	1	91.84% (304)	8.16% (27)
	2	92.92% (197)	7.08% (15)
	3	89.50% (213)	10.50% (25)
	4	77.62% (222)	22.38% (64)
	5	91.88% (362)	8.12% (32)
	6	72.66% (194)	27.34% (73)
	7	84.29% (220)	15.71% (41)
	8	94.44% (306)	5.56% (18)
	9	95.14% (235)	4.86% (12)
	10	75.00% (192)	25.00% (64)
	11	83.38% (291)	16.62% (58)
	12	84.47% (87)	15.53% (16)
	13	83.51% (243)	16.49% (48)
	14	87.40% (222)	12.60% (32)
	15	86.98% (294)	13.02% (44)
	16	95.85% (208)	4.15% (9)
Ethnic identity (Chinese)	Less Chinese	85.59% (1,520)	14.41% (256)
	More Chinese	87.58% (2,270)	12.42% (322)
Proficiency (English)	Less proficient	89.03% (1,193)	10.97% (147)
	More proficient	85.77% (2,597)	14.23% (431)
Gender (speaker)	female	83.80% (1,666)	16.20% (322)
	male	89.24% (2,124)	10.76% (256)
Age	younger	86.52% (1,694)	13.48% (264)
	older	86.97% (2,096)	13.03% (314)
Gender context	mixed	85.73% (1,887)	14.27% (314)
	single	87.82% (1,903)	12.18% (264)
Socioeconomic status or class	lower	84.62% (1,271)	15.38% (231)
	lower middle	89.71% (1,438)	10.29% (165)
	upper middle	85.59% (1,081)	14.41% (182)

Familiarity	more: acquaintances	86.47% (1,112)	13.53% (174)
	less: strangers	86.89% (2,678)	13.11% (404)
Ethnic identity (Hong Konger)	less Hong Konger	90.37% (666)	9.63% (71)
	more Hong Konger	86.04% (3,124)	13.96% (507)
Institution	CUHK	85.85% (1,839)	14.15% (303)
	HKU	87.65% (1,951)	12.35% (275)
Utterance type (Stance)	explanation	90.35% (899)	9.65% (96)
	hard fact	85.84% (297)	14.16% (49)
	opinion	87.83% (1,574)	12.17% (218)
	personal experience	82.59% (1,020)	17.41% (215)
Awareness	less aware	85.22% (1,609)	14.78% (279)
	more aware	87.94% (2,181)	12.06% (299)
Sentiment	negative	90.57% (682)	9.43% (71)
	zero	88.13% (1,463)	11.87% (197)
	positive	84.14% (1,645)	15.86% (310)

C. Topic modelling/ LDA analysis

Table 3. *Results of the LDA analysis: modelled topics, interpretation, and key words.*

Topic	**Interpreted category/topic**	**Top 25 nouns, verbs, and adjectives associated with topic** (in order of probability, highest to lowest)
1	Student Lifestyle and Challenges	*know, cause, school, see, much, want, last, stay, problem, try, year, time, play, sure, need, took, gon, lot, fun, got, person, secondary, interesting, study*
2	Social and Recreational Aspects of School Life	*mean, inaudible, fun, people, money, one, time, play, chance, long, bit, year, food, take, exchange, words, school, get, nice, team, choose, work, think, mathematics, learning*
3	Community and Group Dynamics in Education	*lot, thin, thing, way, know, live, new, society, student, students, okay, going, need, committee, joined, join, bit, used, nice, take, see, years, remember, playing*
4	Academic Environment and Preferences	*different, friends, science, take, many, years, start, write, music, math, wanted, linguistics, hall, want, people, year, form, gon, summer, bit, talk, next, nice, computer*
5	Global and Cultural Exchange in Education	*major, want, people, make, remember, high, travel, okay, thought, top, join, know, nice, year, summer, used, see, take, guess, need, japanese, thinking, talk, playing*
6	Work Experience and Internships	*think, find, sure, job, talk, going, internship, move, depends, time, follow, question, know, bit, local, people, good, need, guess, need, friend, studying, remember, gone*
7	Challenges and Emotional Aspects of Academia	*work, year, difficult, semester, love, bad, week, class, become, people, forgot, let, thinking, guess, second, good, great, team, sort, stuff, research, bit, say, secondary*
8	Graduation and Career Planning	*good, know, time, language, important, graduate, say, need, master, mathematics, shit, come, secondary, nice, think, year, local, wan, summer, read, student, guess, committee, used, studying*
9	Learning Methods and Academic Support	*learn, think, stuff, guess, next, happy, miss, professor, fun, let, play, nice, try, speak, learning, study, bit, good, math, used, say, come, form, guess*
10	University Life and Administration	*got, know, university, hate, year, people, heard, translation, okay, getting, summer, ask, get, speak, secondary, strange, professor, time, took, people, school, studying, thought, say*
11	Extracurricular Activities and Campus Life	*get, chinese, course, archery, life, name, things, give, part, time, hall, said, law, told, school, interesting, call, many, science, one, sort, students, use, person, days*
12	Language Learning and Cultural Studies	*study, see, japanese, say, better, fine, first, minutes, little, day, times, primary, choose, quit, enough, team, play, easy, nice, understand, feel, good, form, need, stuff*

13	Personal Development and Self-Improvement	*know, get, kind, need, friend, studying, final, pay, normal, grade, degree, competition, think, whole, computer, see, enough, good, time, take, okay, playing, bit, sure, people*

D. Quantitative results: Boruta algorithm

Table 4. *Feature importance by variable with high likelihood of conditioning uptalk use: Results of the Boruta algorithm.*

Rank	factor	meanImp	mediamImp	minImp	maxImp	normHits	decision
1	Participant	20.60	20.72	16.93	23.67	1.00	Confirmed
2	Conversation Pair	15.83	15.74	13.26	18.54	1.00	Confirmed
3	Gender of speaker x gender context	12.27	12.16	10.18	14.86	1.00	Confirmed
4	Time	11.71	11.60	9.32	14.00	1.00	Confirmed
5	Utterance type	8.81	8.78	5.61	11.50	1.00	Confirmed
6	Gender context	8.00	8.07	6.07	9.97	1.00	Confirmed
7	Age	7.11	7.21	4.45	9.71	1.00	Confirmed
8	Sentiment	2.50	2.56	-0.26	5.86	0.68	Confirmed

E. Keywords in evaluation experiment

Table 5. *Keywords in Evaluation Experiment by Individual, Gender of Listener, Gender of Speaker, and Uptalk Conditions.*

		Male speaker		**Female speaker**	
		No uptalk	*Uptalk*	*No uptalk*	*Uptalk*
Male Listener	**Samuel**	fluent confident not pretentious not moody	more confident	lower class friendly solidarity less educated Hong Konger positive	more friendly lower class female nicer
	Mason	-	natural normal not excited	natural	hard to tell mood

	Levi	calm rational	confident strong	nerd flat	kind caring
	Johnson	native medical student presentable	non-native leader fluent	Hong Konger emotional effortful caring	not confident normal mood introverted Science student
	George	more confident enunciated	less confident okay English	not emotional	not confident
	Emery	University student Hong Konger slow calm nervous	outgoing talkative	less friendly indifferent robotic	Hong Konger educated not proficient ease
	Tyler	slow	Hong Konger student educated calm	confident brighter smarter clear expressive	dumb
Female listener	**Theodora**	slow emotionless	natural doubtful unsure	female certain carefree	unsure stronger stance
	Tamara	non-Hong Konger confident	uncertain	somewhat uncertain ethnically ambiguous	Hong Konger not confident hesitant
	Hayley	well-travelled strict intelligent not casual	American Educated Casual easygoing	educated calm attentive to detail thoughtful intelligent	Hard to tell education level casual unsure easygoing
	Yasmin	not aggressive debater persuasive	less aggressive educated well-spoken kind	more confident	less confident
	Maya	nervous	confident	nervous	young University middle-class
	Gabriella	confident	hesitant	firm	Uncertain

		certain	uncertain	less uncertain	
	Gia	confident positive	educated younger not mature	less educated negative angry depressed	more confident more positive hesitant higher English proficiency
	Cecilia	clear	optimistic	pessimistic	doubtful

F. Summary table

***Table* 6.** *Classification of mismatching sociolinguistic patterns based on speaker awareness, social evaluation, and usage contexts, with examples.*

Type	Description	Use	Evaluation	Example
latent stereotype	uptalk unconsciously associated with these meanings, but social meanings do not align with relevant patterns of sociolinguistic variation.	-	+ (unaware only)	• Hong Konger
claimed "stereotype"	uptalk judged and claimed to have these meanings regardless of speaker awareness, but social meanings do not align with relevant patterns of sociolinguistic variation.	-	+	• younger • lower/middle class • high familiarity
indicator	uptalk not evaluated with these meanings, but speakers of relevant social groups or speakers placed in relevant contexts use it profusely.	+	-	• mixed-gender • women in mixed-gender settings
marker	uptalk associated with these meanings, but only speakers conditioned in particular contexts and who are less aware of uptalk use it profusely.	+ (unaware only)	+ (regardless of awareness)	• positive