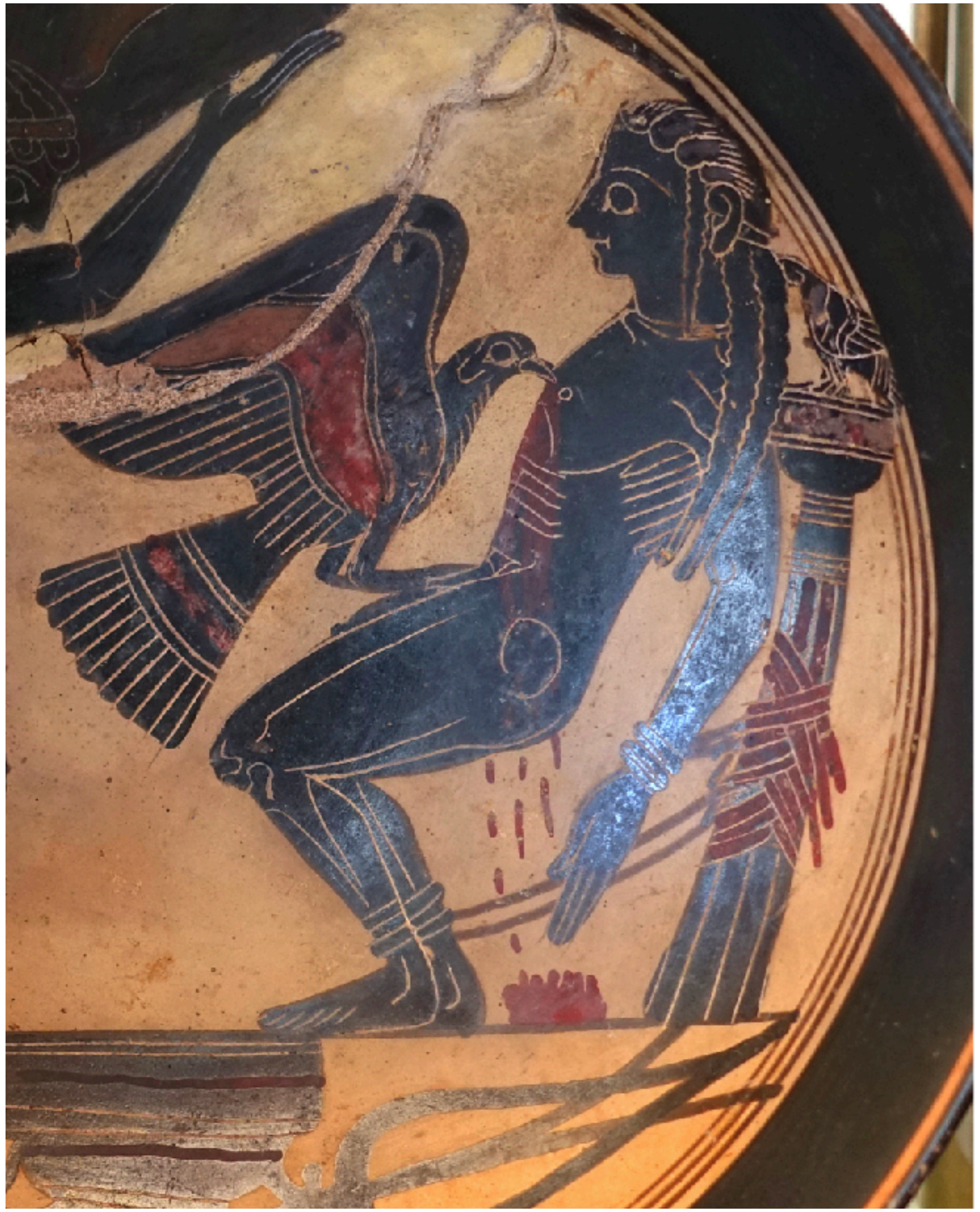
Beyond Individuals: Language in Social Context

CS6120: Natural Language Processing
Northeastern University

David Smith

Prometheus the Fire-Bringer



Prometheus the Fire-Bringer

Still, listen to the miseries that beset mankind—how they were witless before and I made them have sense and endowed them with reason. I will not speak to upbraid mankind but to set forth the friendly purpose that inspired my blessing.

First of all, though they had eyes to see, they saw to no avail; they had ears, but they did not understand ; but, just as shapes in dreams, throughout their length of days, without purpose they wrought all things in confusion. They had neither knowledge of houses built of bricks and turned to face the sun nor yet of work in wood; but dwelt beneath the ground like swarming ants, in sunless caves. They had no sign either of winter or of flowery spring or of fruitful summer, on which they could depend but managed everything without judgment, until I taught them to discern the risings of the stars and their settings, which are difficult to distinguish.

Yes, and numbers, too, chiefest of sciences, I invented for them, and the combining of letters, creative mother of the Muses' arts, with which to hold all things in memory.

Aeschylus *Prometheus Bound* 442–461

The Secret of Our Success

Physically weak, slow… dependent on eating cooked food, though we don't innately know how to make fire or cook…. Our colons are too short, stomachs too small, and teeth too petite. Our infants are born fat and dangerously premature… not so impressive when we go head-to-head in problem-solving tests against other apes…. We are a **cultural** species. Probably over a million years ago, members of our evolutionary lineage began learning from each other in such a way that culture became cumulative…. *Our capacities for learning from others are themselves finely honed products of natural selection*…. Cultural learning abilities gave rise to an interaction between an accumulating body of cultural information and genetic evolution that has shaped, and continues to shape, our anatomy, physiology, and psychology…

Henrich (2016), emphasis mine

Cultural Technologies (Farrell et al. 2025)

- LLMs are the latest **cultural technologies** that allow humans to get knowledge from, and coordinate with, humans in other times and places
- Most obvious comparisons are language, writing, print, libraries, the internet
- Less obvious are markets, democracies, and bureaucracies

Cultural Technologies (Farrell et al. 2025)

- Markets, democracies, and bureaucracies (Weber, Hayek, and all that) aggregate and summarize complicated information across societies into prices, votes, and procedures
- Different dynamics of homogenization and fragmentation
- These institutions (need to) attenuate the diversity of languages in circulation (Gellner 1983), like LLMs

Cultural Technologies

- Many advantages from LLMs being trained on more data than a single human agent could produce/consume
- Language, itself a cultural technology, operates to connect human agents
- How does language reflect human identities, relationships, and power?
- How do human identities, relationships, and power—as mediated through language—affect AI models?

On the internet, no one knows you're a …

- Mosteller & Wallace 1963. Inference in an authorship problem.
  - *Only* use common stopwords to avoid topic confounding
- Advertising demands demographic prediction
  - Why? Why not just predict conversions?
- Sarawgi et al. 2011, Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre
  - Balance topics to avoid overestimating accuracy

On the internet, no one knows you're a …

	lexicon based		deep syntax	morphology			b.o.w.	shallow lex-syntax		
Topic	Gender Genie	Gender Guesser	PCFG	CLM n=1	CLM n=2	CLM n=3	ME	TLM n=1	TLM n=2	TLM n=3
Per Topic Accuracy (%) for All Authors										
Entertain	50.0	42.5	50.0	52.5	**67.5**	**67.5**	60.0	57.5	57.5	57.5
Book	50.0	42.5	65.0	57.5	67.5	**72.5**	55.0	60.0	67.5	67.5
Politics	35.0	30.0	50.0	47.5	**52.5**	50.0	45.0	**52.5**	**52.5**	**52.5**
History	40.0	35.0	77.5	65.0	**80.0**	**80.0**	55.0	65.0	65.0	65.0
Education	62.5	42.5	55.0	63.0	65.0	**70.0**	63.0	55.0	57.5	52.5
Travel	62.5	37.5	63.0	**65.0**	63.0	63.0	63.0	62.5	**65.0**	**65.0**
Spirituality	50.0	32.5	53.0	**78.0**	**78.0**	**78.0**	50.0	65.0	70.0	72.5
Avg	50.0	37.5	59.0	61.2	**68.3**	**68.3**	55.87	60.0	61.3	61.5
Per Topic Accuracy (%) for Female Authors										
Entertain	25.0	10.0	**85.0**	70.0	50.0	**85.0**	70.0	75.0	75.0	75.0
Book	15.0	15.0	**95.0**	80.0	**95.0**	90.0	85.0	75.0	90.0	90.0
Politics	10.0	05.0	**65.0**	00.0	05.0	00.0	35.0	30.0	30.0	25.0
History	10.0	05.0	**90.0**	70.0	80.0	75.0	70.0	50.0	50.0	50.0
Education	45.0	10.0	80.0	95.0	85.0	90.0	**100.0**	50.0	55.0	50.0
Travel	65.0	00.0	85.0	90.0	**100.0**	**100.0**	**100.0**	85.0	95.0	90.0
Spirituality	20.0	00.0	60.0	65.0	65.0	**70.0**	45.0	50.0	50.0	50.0
Avg	27.1	06.4	**80.0**	67.1	68.6	72.9	72.1	59.3	63.6	61.4
Per Topic Accuracy (%) for Male Authors										
Entertain	75.0	75.0	15.0	35.0	**85.0**	50.0	50.0	40.0	40.0	40.0
Book	**80.0**	70.0	35.0	35.0	40.0	55.0	25.0	45.0	45.0	45.0
Politics	60.0	55.0	35.0	95.0	**100.0**	**100.0**	55.0	75.0	75.0	80.0
History	70.0	65.0	65.0	60.0	80.0	**85.0**	40.0	80.0	80.0	80.0
Education	**80.0**	75.0	30.0	30.0	45.0	50.0	25.0	60.0	60.0	55.0
Travel	60.0	**75.0**	40.0	40.0	25.0	25.0	25.0	40.0	35.0	40.0
Spirituality	80.0	65.0	45.0	**90.0**	**90.0**	85.0	55.0	80.0	90.0	95.0
Avg	**72.1**	68.6	37.9	55.0	66.4	64.2	39.3	60.0	60.8	62.1

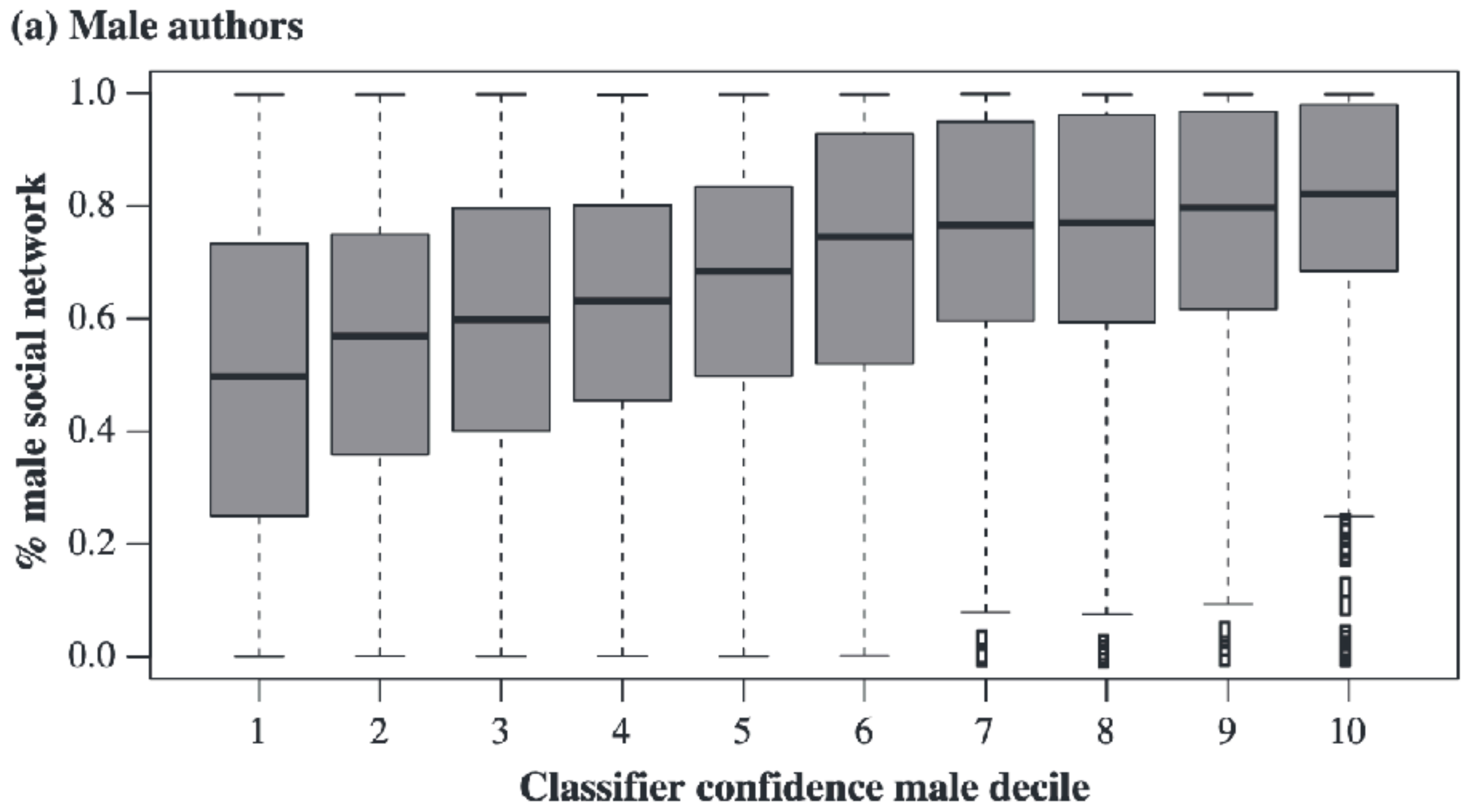
Gender and homophily

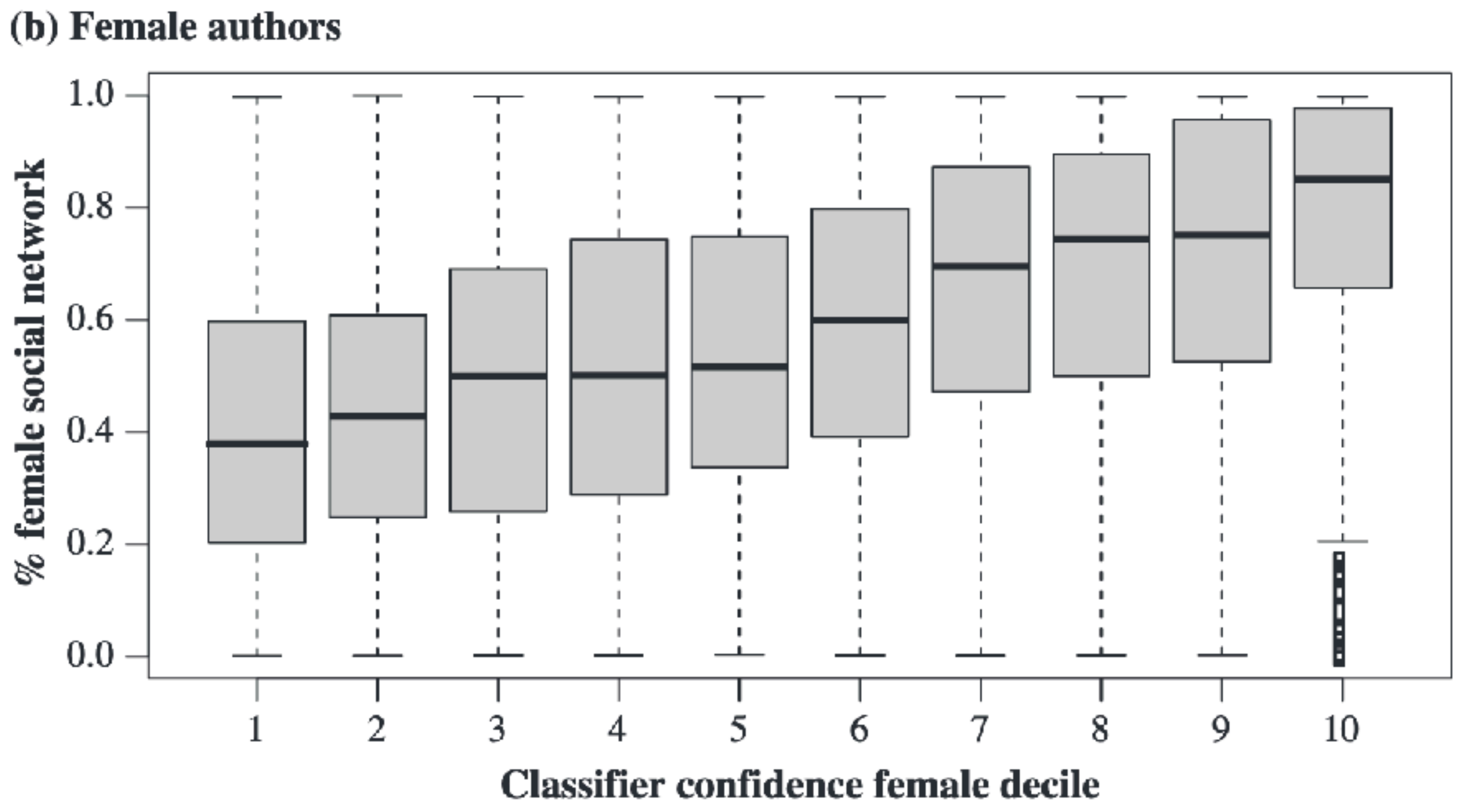
- Gender is a *performance*
- More and less mixed environments license different kinds of language from the same people
- Bamman et al. 2014. Gender identity and lexical variation in social media

Table 1: Comparison of gender markers with previous research ('ns' indicates no significant association; 'mixed' indicates markers for male and female genders)

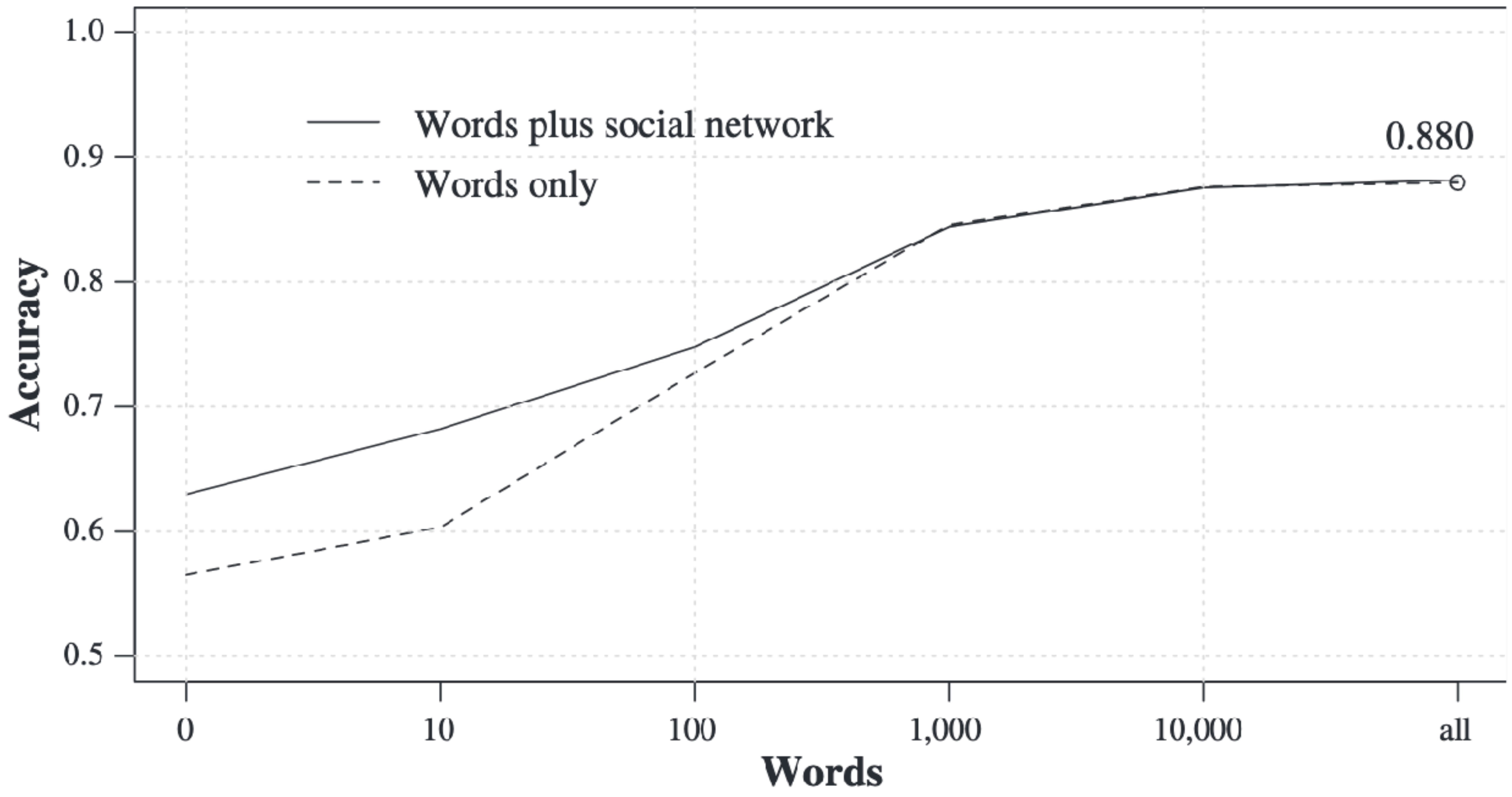
Word class	Previous literature	Our analysis
Pronouns	F	F
Emotion terms	F	F
Kinship terms	F	mixed
CMC words (*lol*, *omg*)	F	F
Conjunctions	F	ns
Clitics	F	ns
Articles	M	ns
Numbers	M	M
Quantifiers	M	ns
Technology words	M	M
Prepositions	mixed	ns
Swear words	mixed	M
Assent	mixed	F
Negation	mixed	mixed
Emoticons	mixed	F
Hesitation	mixed	F

Gender and homophily





Gender and homophily

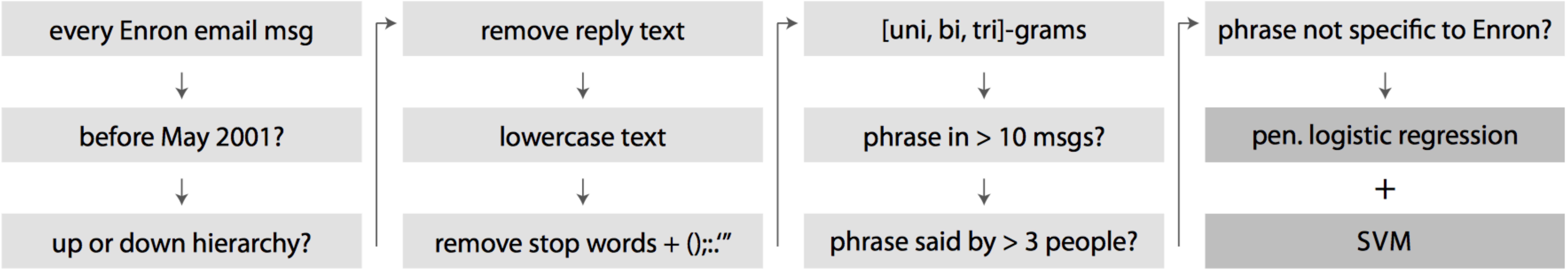


Unequal relationships

- Language use is relational
- Human relationships are mediated by power
- Gilbert 2012. Phrases that signal workplace hierarchy.
  - Emails from the Enron corporation, released as part of a fraud investigation in 2001 and a staple of NLP ever since
  - Predict Enron org chart: for each sender/recipient(s), who is higher in the organization?

Unequal relationships

- Bag of n-grams document representation
- Binary classification



Unequal relationships

these predict message going up

not going up

↑ phrases	β	↑ phrases	β	↔↓ phrases	β	↔↓ phrases	β
the ability to	6.76	attach	6.72	have you been	-8.46	to manage the	-6.66
I took	6.57	that we might	6.54	you gave	-6.64	let's discuss	-5.72
are available	6.52	the calendar	6.06	we are in	-5.44	publicly	-5.24
kitchen	5.72	can you get	5.72	title	-5.05	promotion	-5.02
thought you would	5.65	driving	5.61	need in	-4.80	good one	-4.62
, I'll be	5.51	thoughts on	5.51	opened	-4.57	determine the	-4.47
looks fine	5.50	shit	5.45	initiatives	-4.38	is difficult	-4.36
voicemail	5.43	we can talk	5.41	. I would	-4.34	man	-4.26
tremendous	5.27	it does	5.21	we will probably	-4.12	number we	-4.11
will you	5.17	involving	5.15	any comments	-4.06	contact you	-4.05
left a	5.07	the report	5.04	you said	-3.99	the problem is	-3.97
I put	4.90	please change	4.88	I left	-3.88	you did	-3.78
you ever	4.80	issues I	4.76	can you help	-3.68	cool	-3.54
I'll give	4.69	is really	4.65	send this	-3.47	your attention	-3.44
okay ,	4.60	your review	4.56	whether we	-3.44	to think	-3.44
to send it	4.48	europe	4.45	the trade	-3.40	addition to the	-3.30
communications	4.38	weekend .	4.35	and I thought	-3.28	great thanks	-3.24
a message	4.35	have our	4.33	should include	-3.19	selected	-3.16
one I	4.28	interviews	4.28	please send	-3.14	ext	-3.13
can I get	4.28	you mean	4.26	existing	-3.06	and let me	-3.05
worksheet	4.15	haven't been	4.10	mondays	-3.02	security	-3.01
liked	4.07	me . 1	4.07	presentation on	-2.95	got the	-2.94
I gave you	3.95	tiger	3.94	let's talk	-2.94	get your	-2.88
credit will	3.88	change in	3.88	the items	-2.78	this week and	-2.77
you make	3.86	item	3.84	i hope you	-2.77	team that	-2.75
together and	3.82	a decision	3.82	did it	-2.75	a deal	-2.71
have presented	3.78	a discussion	3.74	test	-2.69	yours .	-2.68
think about	3.71	sounds good	3.65	be sure	-2.65	briefing	-2.60

Accommodating power

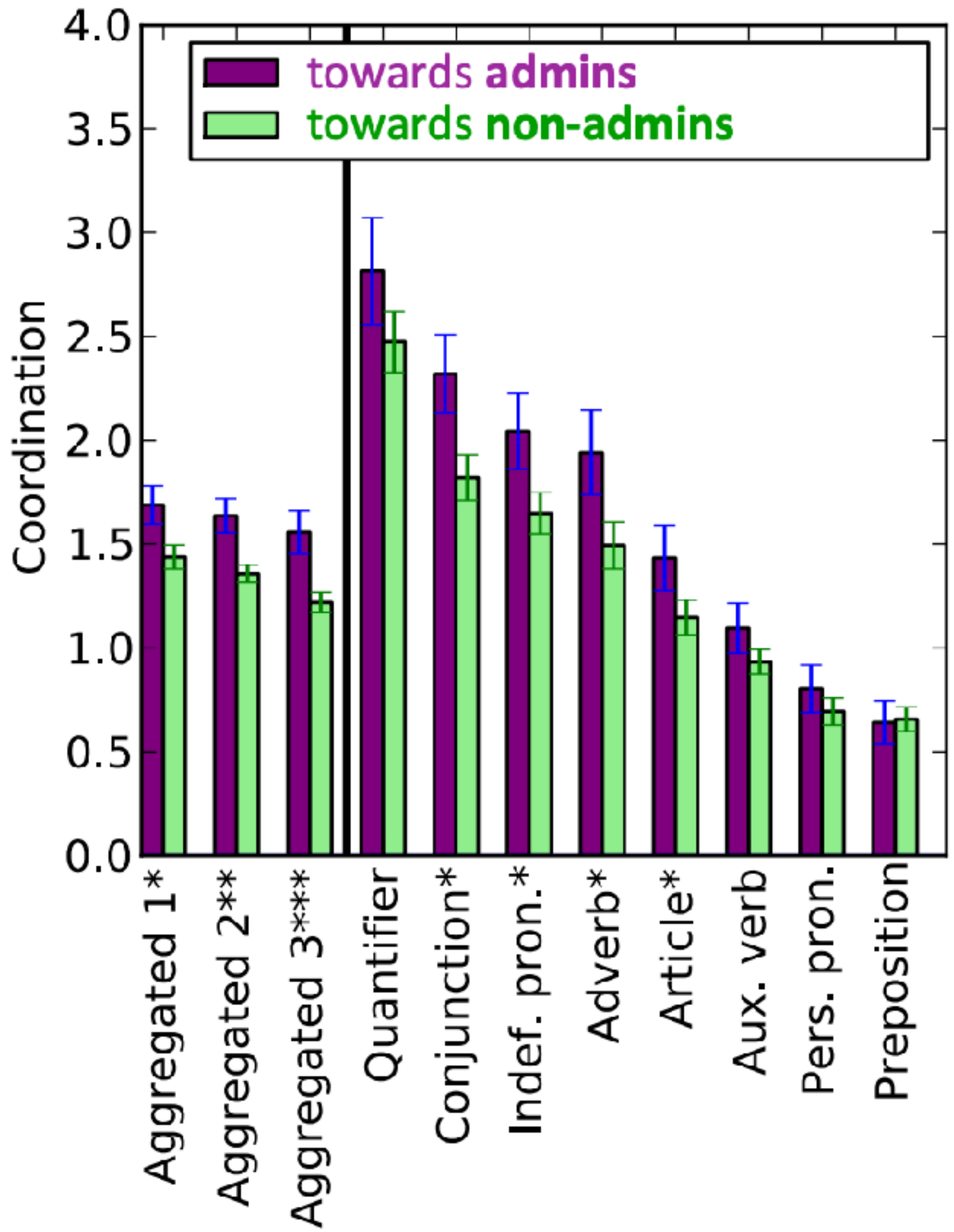
- Danescu-Niculescu-Mizil et al. 2012. Echoes of Power: Language Effects and Power Differences in Social Interaction.
  - Given Wikipedia discussion pages, or US Supreme Court oral arguments
  - Differing status:
    - Wikipedia admin/non-admin
    - SCOTUS justices/lawyers
  - Measure linguistic **coordination**: Adapting language to higher-status speakers

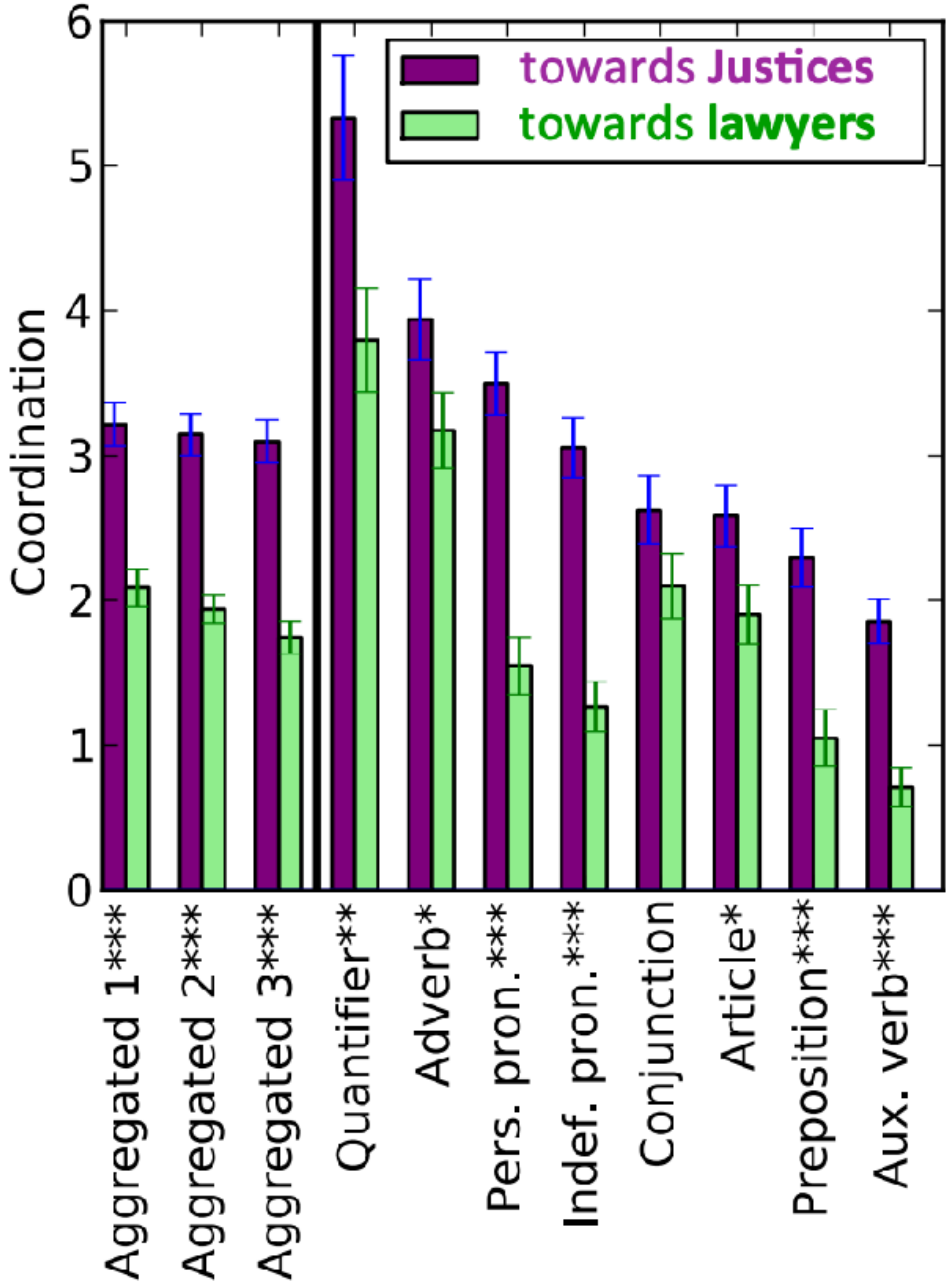
Accommodating power

- Danescu-Niculescu-Mizil et al. 2012. Echoes of Power: Language Effects and Power Differences in Social Interaction.
  - Measure linguistic **coordination**: Adapting language to higher-status speakers

$$Coordination_{(J\ to\ V)}(art.) = P(J^{art.} \mid J \text{ replied to } V, \overbrace{V^{art}}^{\text{Trigger}}) - \underbrace{P(J^{art.} \mid J \text{ replied to } V)}_{\text{Control (for inherent similarity)}}$$

Accommodating power





Measuring respect

- Outside fixed hierarchies, human relations are still mediated by power and respect
- Voigt et al. 2017. Language from police body camera footage shows racial disparities in officer respect.
  - Transcripts of 981 Oakland, CA, traffic stops
  - Rate one police/driver turn on 4-point scale, high inter-annotator agreement

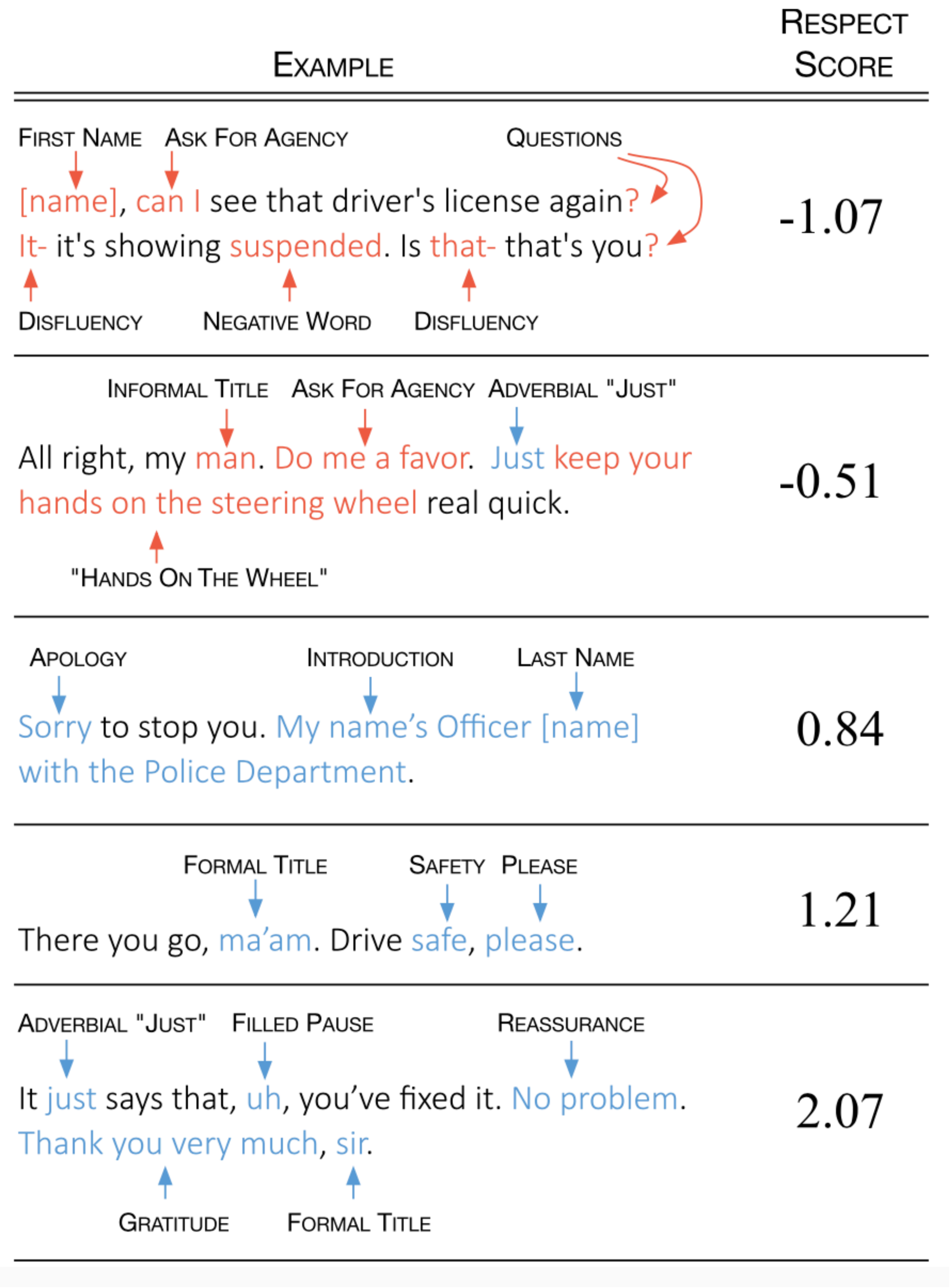
Measuring respect

Feature Name	Implementation
Adverbial "Just"	"Just" occurs in a dependency arc as the head of an `advmod` relation
Apologizing	Lexicon: "sorry", "oops", "woops", "excuse me", "forgive me", "apologies", "apologize", "my bad", "my fault"
Ask for Agency	Lexicon: "do me a favor", "let me", "allow me", "can i", "should i", "may i", "might i", "could i"
Bald Command	The first word in a sentence is a bare verb with part-of-speech tag VB ("look", "give", "wait" etc.) but is not one of "be", "do", "have", "thank", "please", "hang".
Colloquialism	Regular expression capturing "y'all", "ain't" and words ending in "in'" such as "walkin'", "talkin'", etc., as marked by transcribers
Conditional	Lexicon: "if"
Disfluency	Word fragment ("Well I thi-") as indicated by transcribers
Filled Pauses	Lexicon: "um", "uh"
First Names	Top 1000 most common first names from the 1990 US Census, where first letter is capitalized in transcript
Formal Titles	Lexicon: "sir", "ma'am", "maam", "mister", "mr*", "ms*", "madam", "miss", "gentleman", "lady"
For Me	Lexicon: "for me"
For You	Lexicon: "for you"
Give Agency	Lexicon: "let you", "allow you", "you can", "you may", "you could"
Gratitude	Lexicon: "thank", "thanks", "appreciate"
Goodbye	Lexicon: "goodbye", "bye", "see you later"
Hands on the Wheel	Regular expression capturing cases like "keep your hands on the wheel" and "leave your hands where I can see them": `"hands? ([÷,?!:;]+ )?(wheel|see)"`

Measuring respect

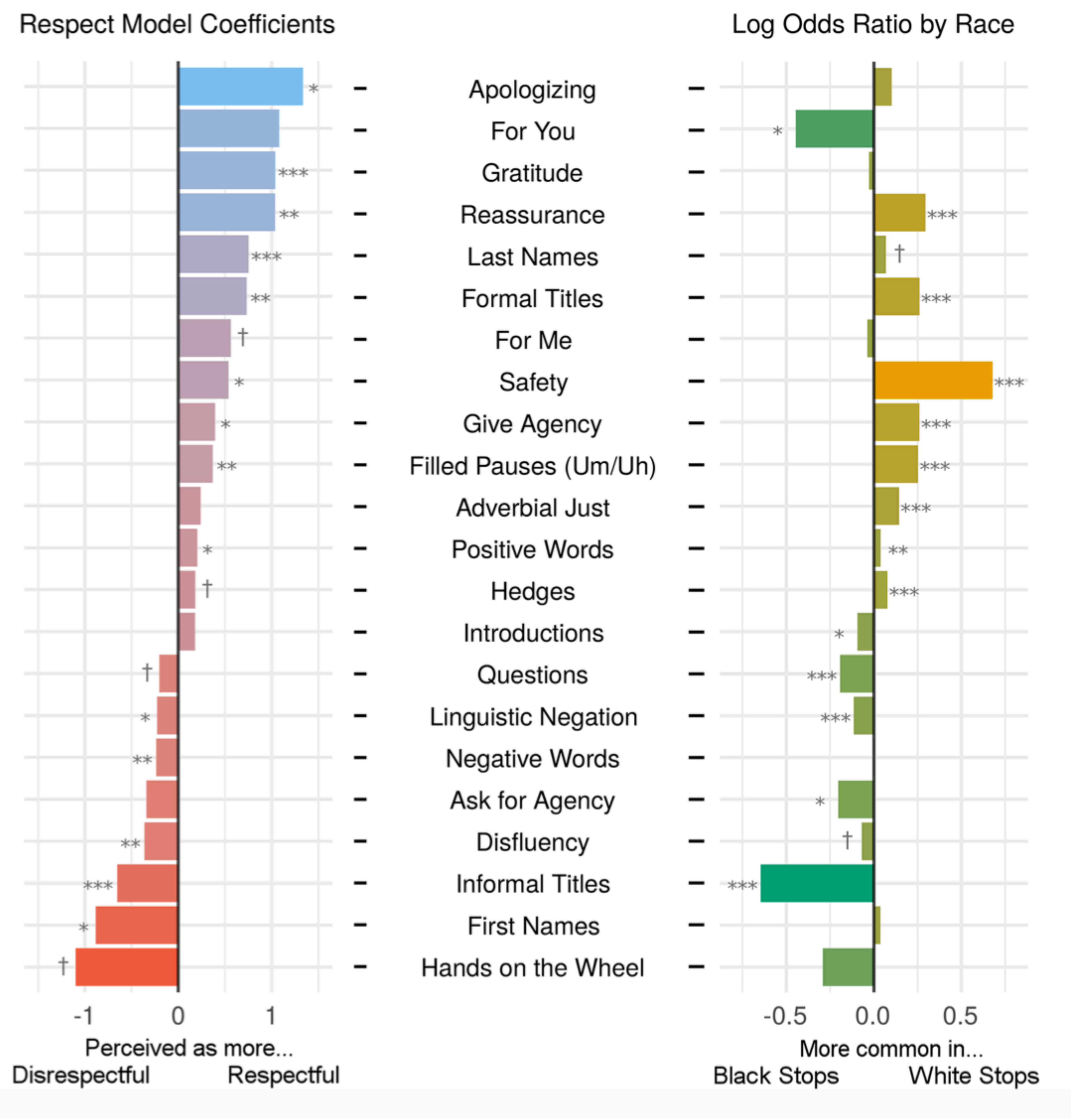
<table>
<tr><td>Hedges</td><td>All words in the "Tentat" LIWC lexicon</td></tr>
<tr><td>Impersonal Pronoun</td><td>All words in the "Imppron" LIWC lexicon</td></tr>
<tr><td>Informal Titles</td><td>Lexicon: "dude*", "bro*", "boss", "bud", "buddy", "champ", "man", "guy*", "guy", "brotha", "sista", "son", "sonny", "chief"</td></tr>
<tr><td>Introductions</td><td>Regular expression capturing cases like "I'm Officer [name] from the OPD" and "How's it going?": "((i|my name).+officer | officer.+(oakland|opd))|((hi|hello|hey|good afternoon|good morning|good evening|how are you doing|how 's it going))"</td></tr>
<tr><td>Last Names</td><td>Top 5000 most common last names from the 1990 US Census, where first letter is capitalized in transcript</td></tr>
<tr><td>Linguistic Negation</td><td>All words in the "Negate" LIWC lexicon</td></tr>
<tr><td>Negative Words</td><td>All words in the "Negativ" category in the Harvard General Inquierer, matching on word lemmas</td></tr>
<tr><td>Positive Words</td><td>All words in the "Positiv" category in the Harvard General Inquierer, matching on word lemmas</td></tr>
<tr><td>Please</td><td>Lexicon: "please"</td></tr>
<tr><td>Questions</td><td>Occurrence of a question mark</td></tr>
<tr><td>Reassurance</td><td>Lexicon: "'s okay", "n't worry", "no big deal", "no problem", "no worries", "'s fine", "you 're good", "is fine", "is okay"</td></tr>
<tr><td>Safety</td><td>Regular expression for all words beginning with the prefix "safe", such as "safe", "safety", "safely"</td></tr>
<tr><td>Swear Words</td><td>All words in the "Swear" LIWC lexicon</td></tr>
<tr><td>Tag Question</td><td>Regular expression capturing cases like "..., right?" and "..., don't you?": ", (((all right|right|okay|yeah|please|you know)(sir| ma'am| miss| son)?)|((are|is|do|can|have|will|won't) (n't)?(i|me|she|us|we|you|he|they|them))) [?]"</td></tr>
<tr><td>The Reason for the Stop</td><td>Lexicon: "reason", "stop* you", "pull* you", "why i", "why we", "explain", "so you understand"</td></tr>
<tr><td>Time Minimizing</td><td>Regular expression capturing cases like "in a minute" and "let's get this done quick": "(a|one|a few) (minute|min|second|sec|moment)s?|this[.,?!]+quick|right back"</td></tr>
</table>

Measuring respect



Measuring respect

- Higher respect to white drivers, older drivers, when a citation is issued
- Lower respect when a search is conducted

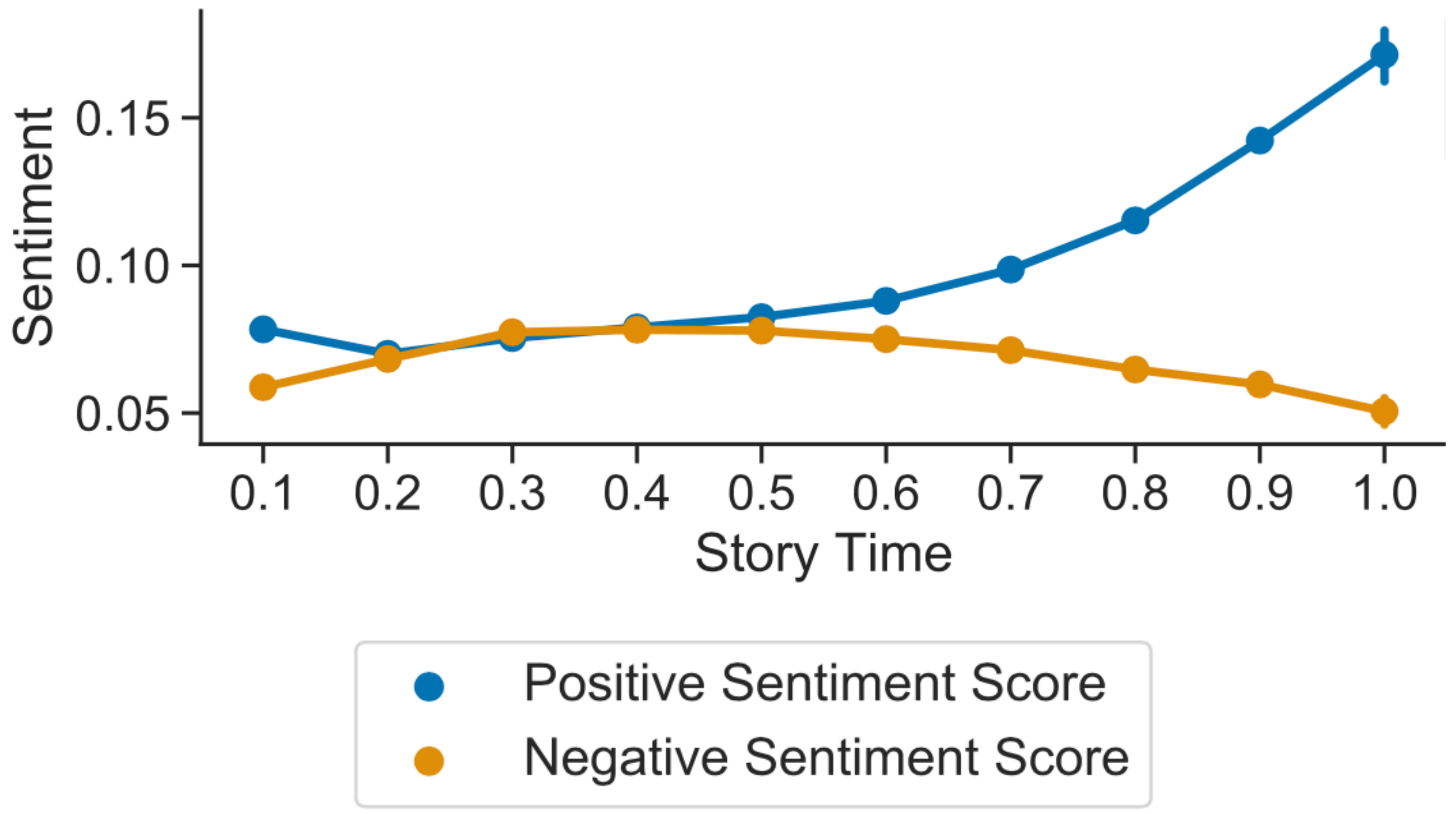


Measuring agency in birth stories

- Antoniak et al. 2019. Narrative paths and negotiation of power in birth stories.
  - 2,847 birth stories from r/BabyBumps — "narratives of individual experiences giving birth, often in great medical and emotional detail"
  - Analyzing narrative arcs with
    - Topic modeling (unigram LMs clustering tokens in documents into coherent "topics")
    - Sentiment analysis
    - Connotation frames of power

Narrative arcs in birth stories

- Dictionary-based sentiment analysis with VADER lexicon (Hutto and Gilbert 2014)



Topic models

- Input: documents, number of topics
- Output:
  - **Unigram distribution for each topic**
  - Topic distribution for each document
  - Topic distribution for each word

{album, band, music}	{government, party, election}	{game, team, player}
album	government	game
band	party	team
music	election	player
song	state	win
release	political	play

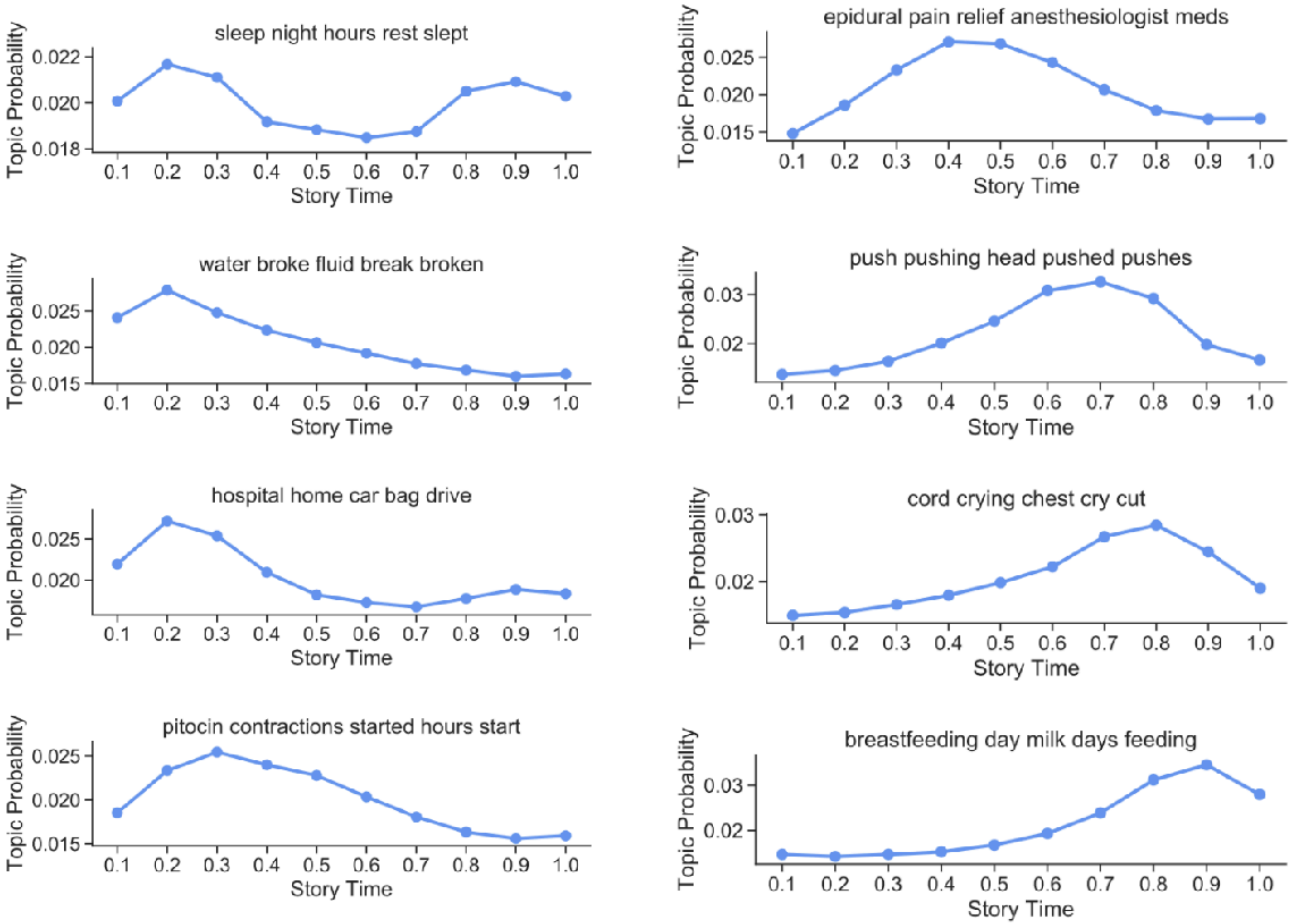
{god, call, give}	{company, market, business}	{math, number, function}
god	company	math
call	market	number
give	business	function
man	year	code
time	product	set

{city, large, area}	{math, energy, light}	{law, state, case}
city	math	law
large	energy	state
area	light	case
station	field	court
include	star	legal

Topic models for birth stories

- Run Latent Dirichlet Allocation (LDA) on training birth stories, each divided into 100-word chunks, with 50 topics
- Divide each story into 10 chunks, plot aggregate topic distribution over narrative time.

Topic models for birth stories



Personas for narrative actors

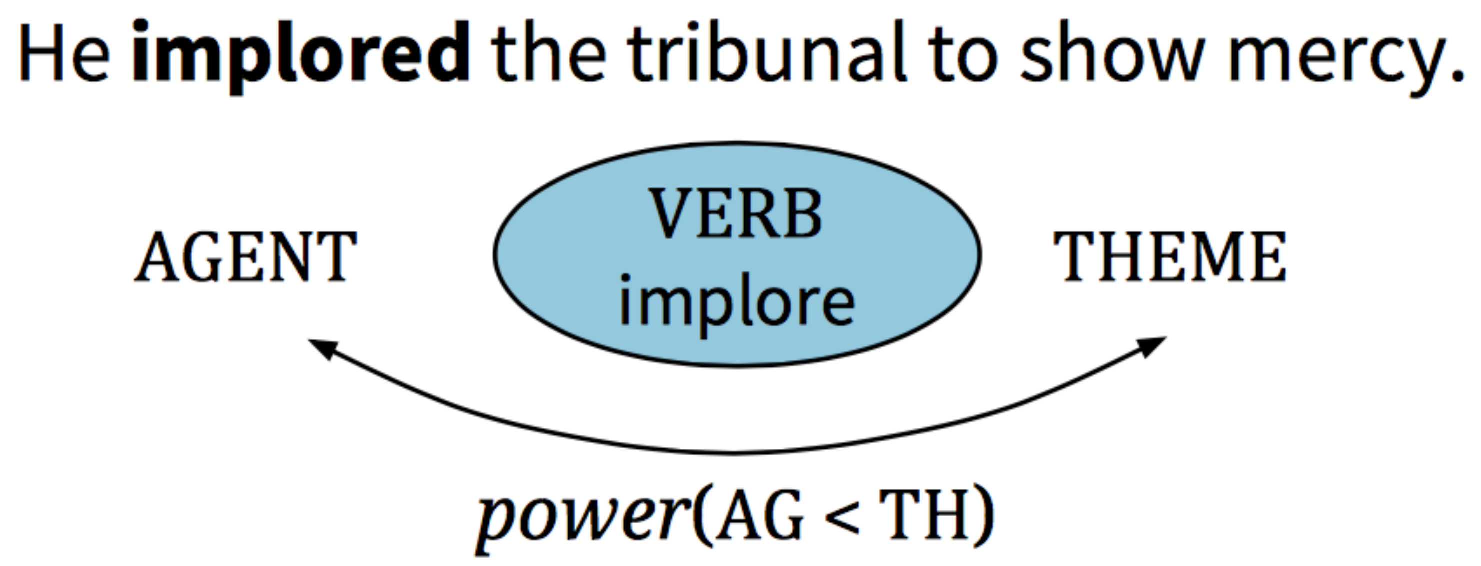
- Dictionary-based method to group word types into "personas" — e.g., partner, husband, wife → PARTNER

Persona	N-Grams	Total Mentions	Stories Containing Mentions	Average Mentions per Story
AUTHOR	I, me, myself	210,795	2,846	74.0
We	we, us, ourselves	24,757	2,764	8.7
BABY	baby, son, daughter	14,309	2,668	5.0
DOCTOR	doctor, dr, doc, ob, obgyn, gynecologist, physician	10,025	2,262	3.5
PARTNER	partner, husband, wife	8,998	2,006	3.2
NURSE	nurse	7,080	2,012	2.5
MIDWIFE	midwife	4,069	886	1.4
FAMILY	mom, dad, mother, father, brother, sister	3,490	1,365	1.2
ANESTHESIOLOGIST	anesthesiologist	1,398	876	0.5
DOULA	doula	896	256	0.3

Table 5. Personas identified in the birth stories collection and the n-grams used to classify the personas.

Power frames

- Sap et al. 2017 Connotation Frames of Power and Agency in Modern Films: Verbs imply power differential between agent/theme



$power(\text{AG}<\text{TH})$

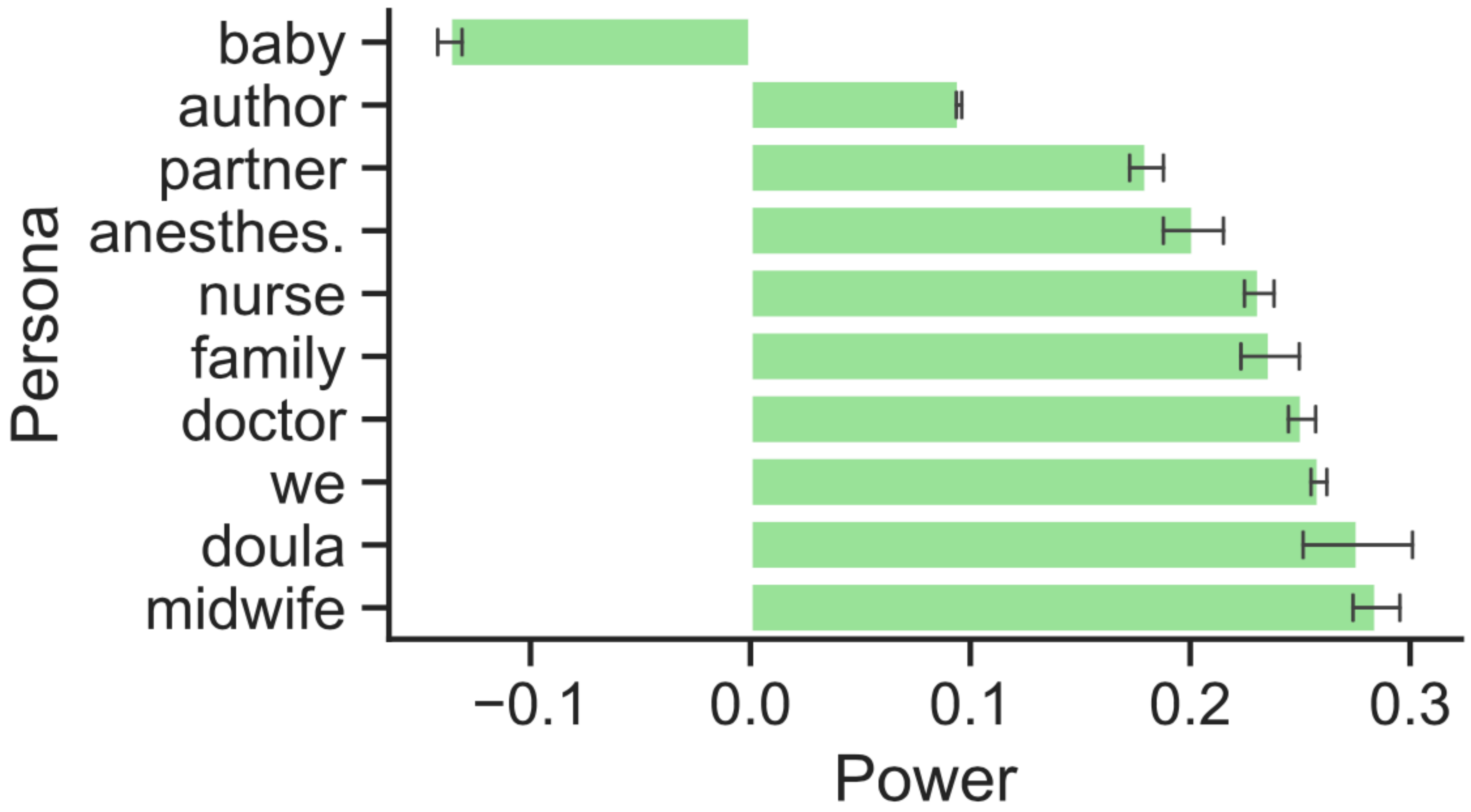
$power(\text{AG}>\text{TH})$

Power frames in birth stories

- The only time I got upset was when the **nurse** accused *me* of not feeding my child.
- The **doctor** broke my water.

Power frames in birth stories

- The author is framed as having the **least power** (except for the baby).
- Clinicians are framed as having high power



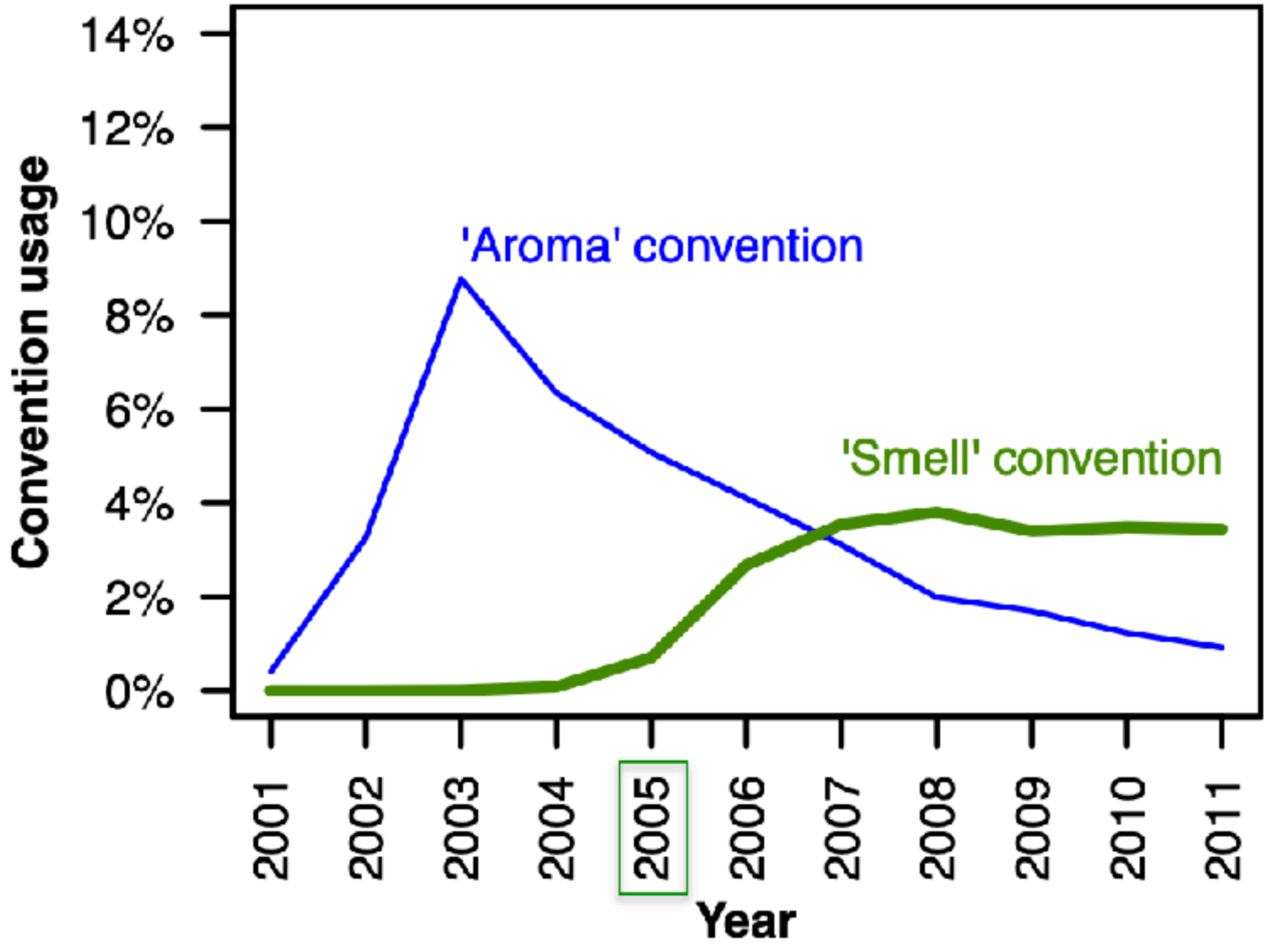
Changing language as in-group signaling

- Danescu et al. 2013. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities.
  - Beer Advocate: 10y, 1.6M posts, 33k users
  - rateBeer: 10y, 3M posts, 30k users
  - How does the community's language change?
  - How does users' language change over time in the community?

Changing language as in-group signaling

… Aroma: Buttery, slightly spicy malt notes …

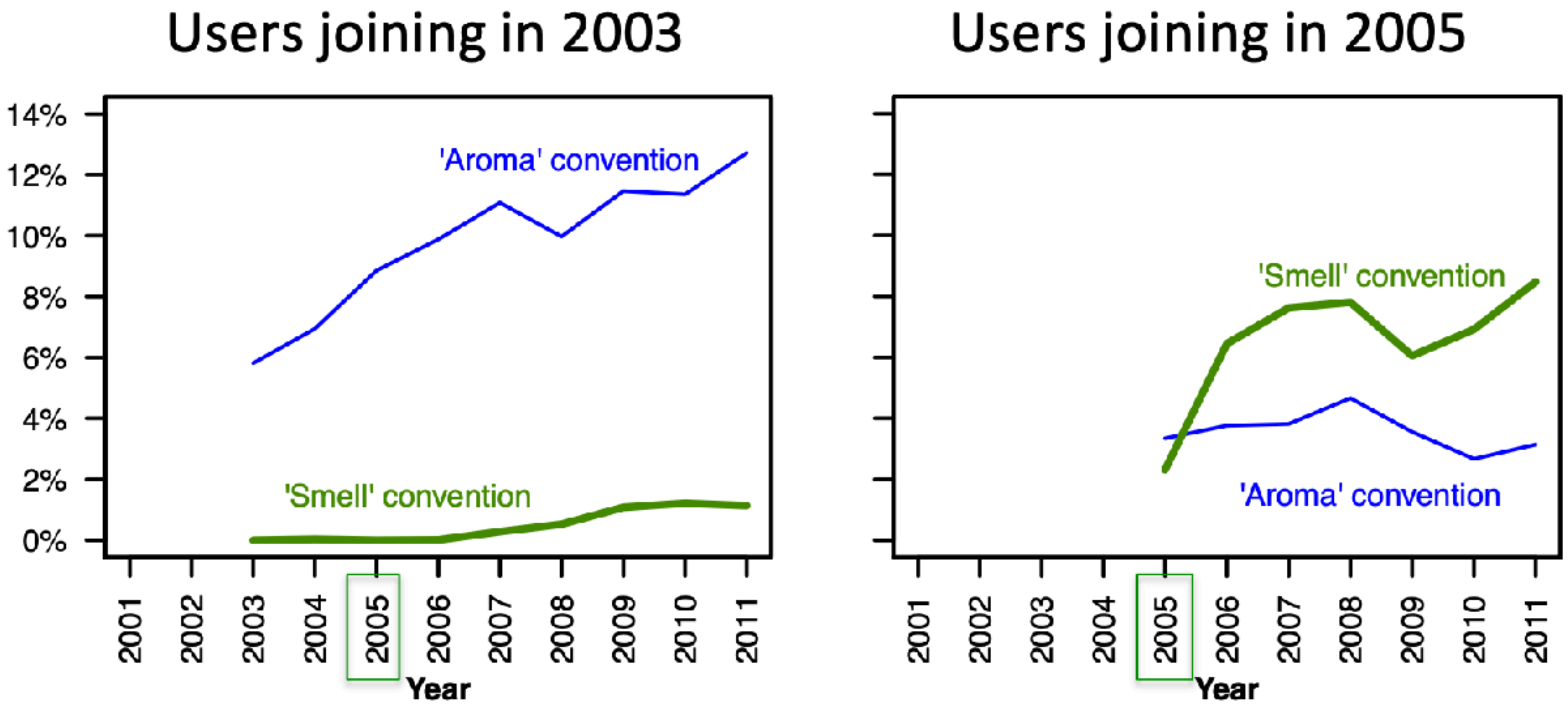
… S: Great nose of ginger, honey, perfume …



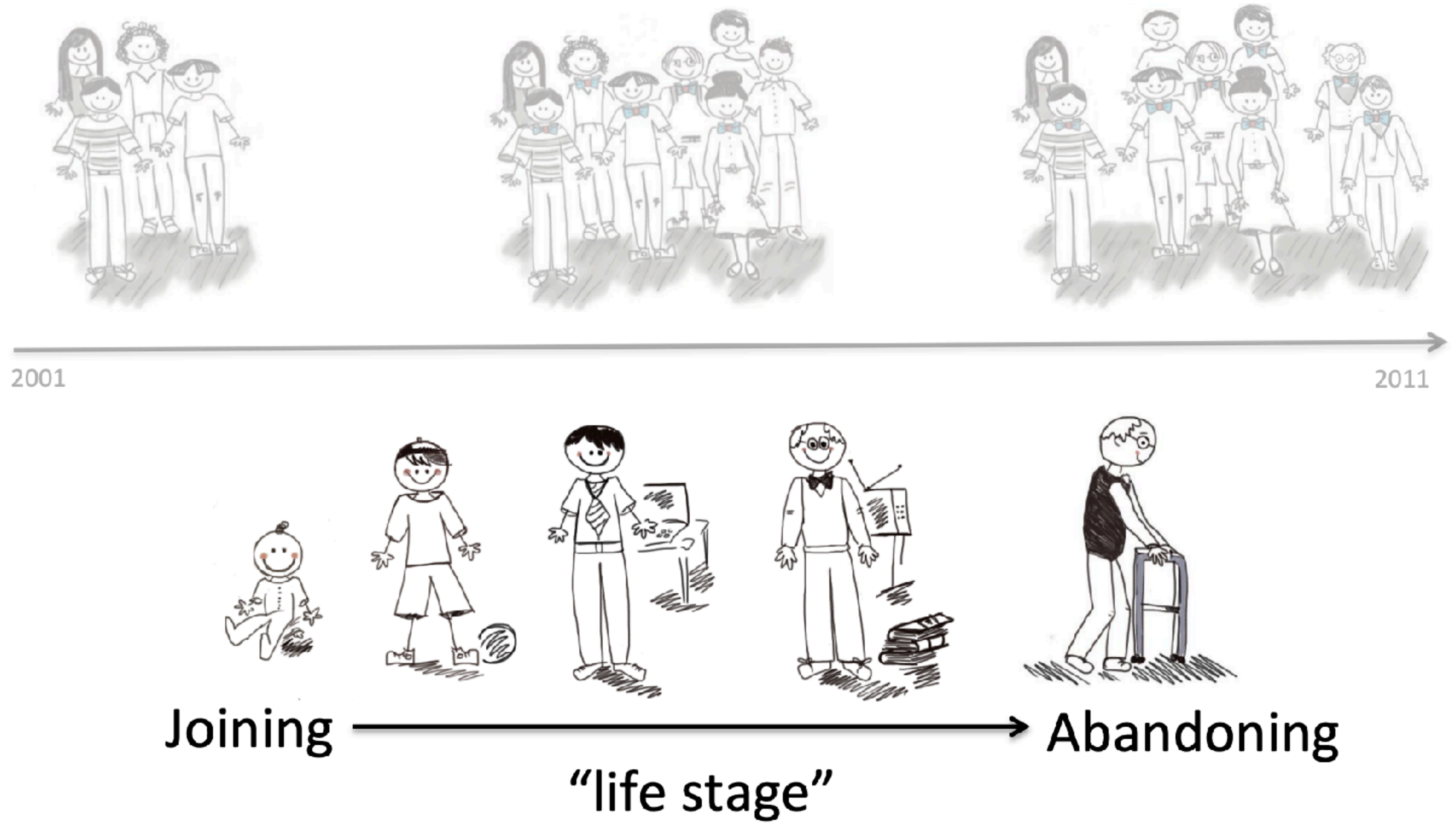
Changing language as in-group signaling

... Aroma: Buttery, slightly spicy malt notes ...

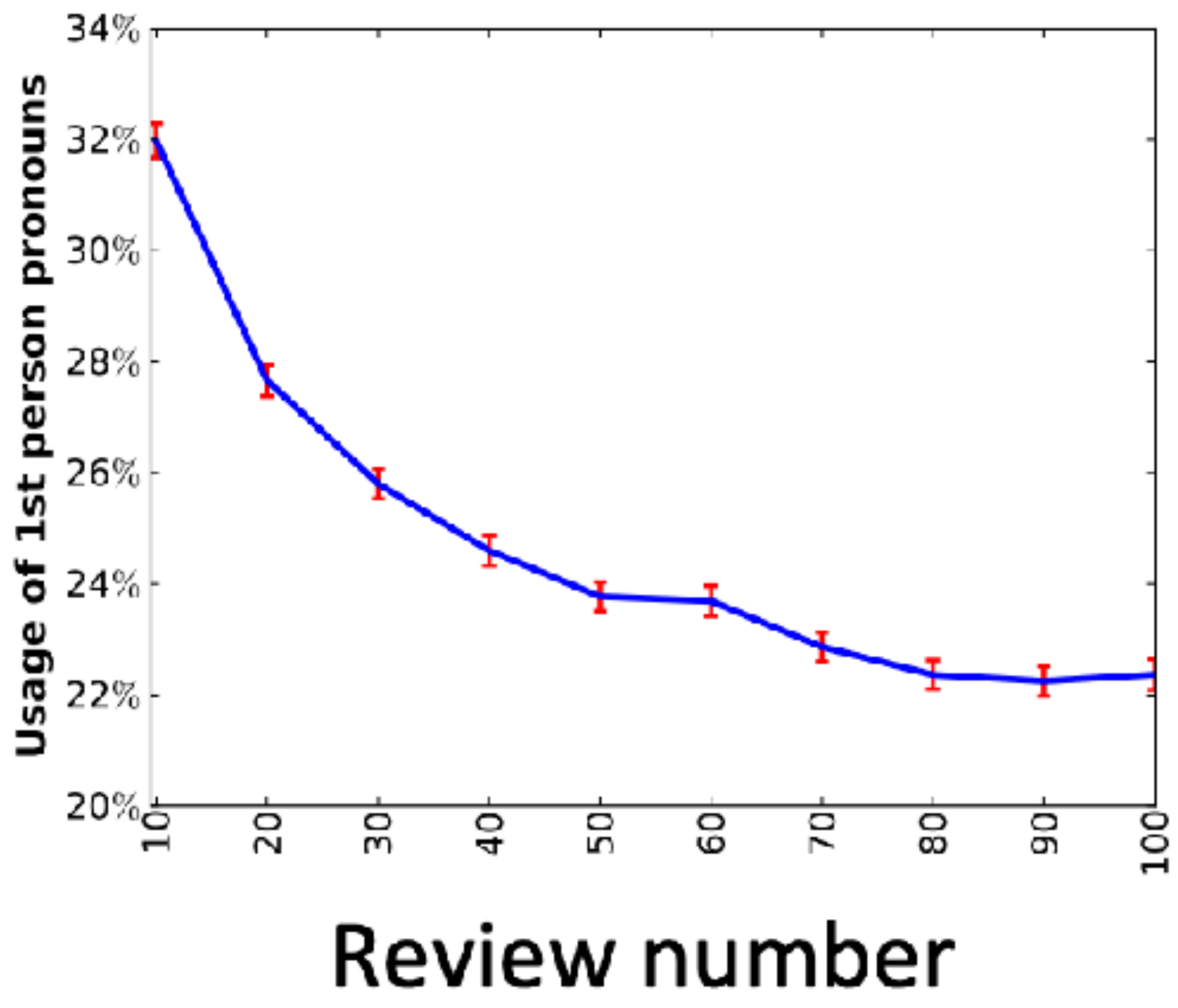
... S: Great nose of ginger, honey, perfume ...



Community- and user-level changes

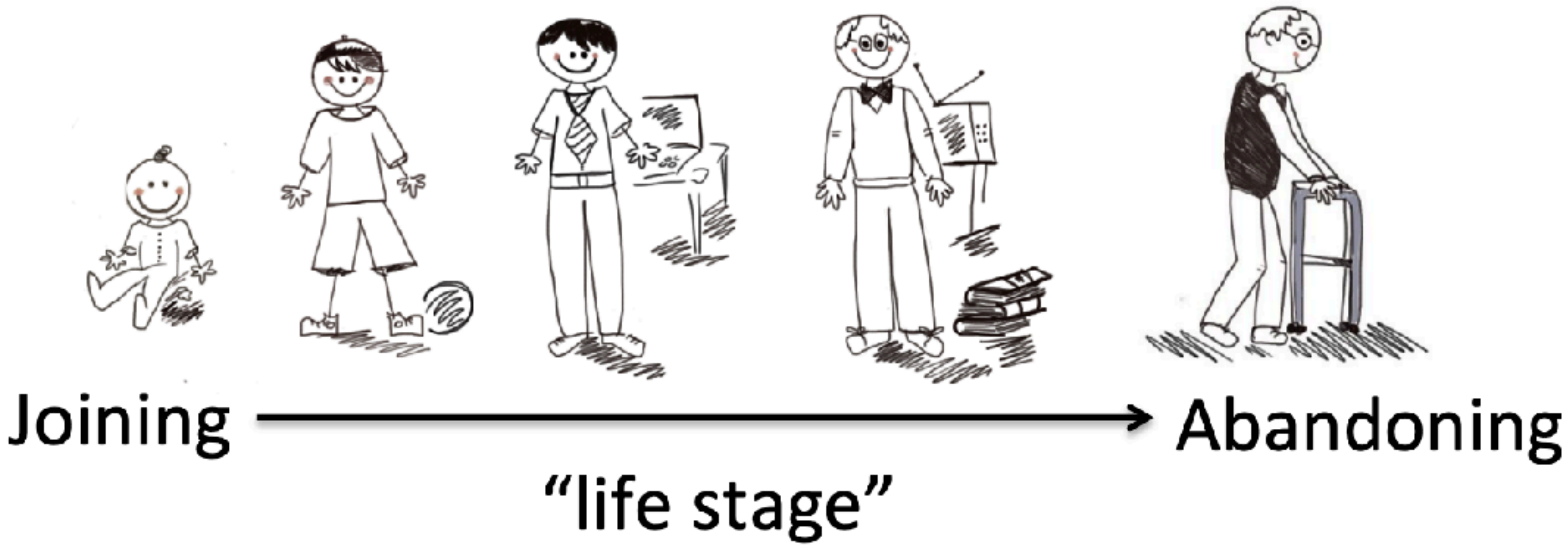


Community- and user-level changes

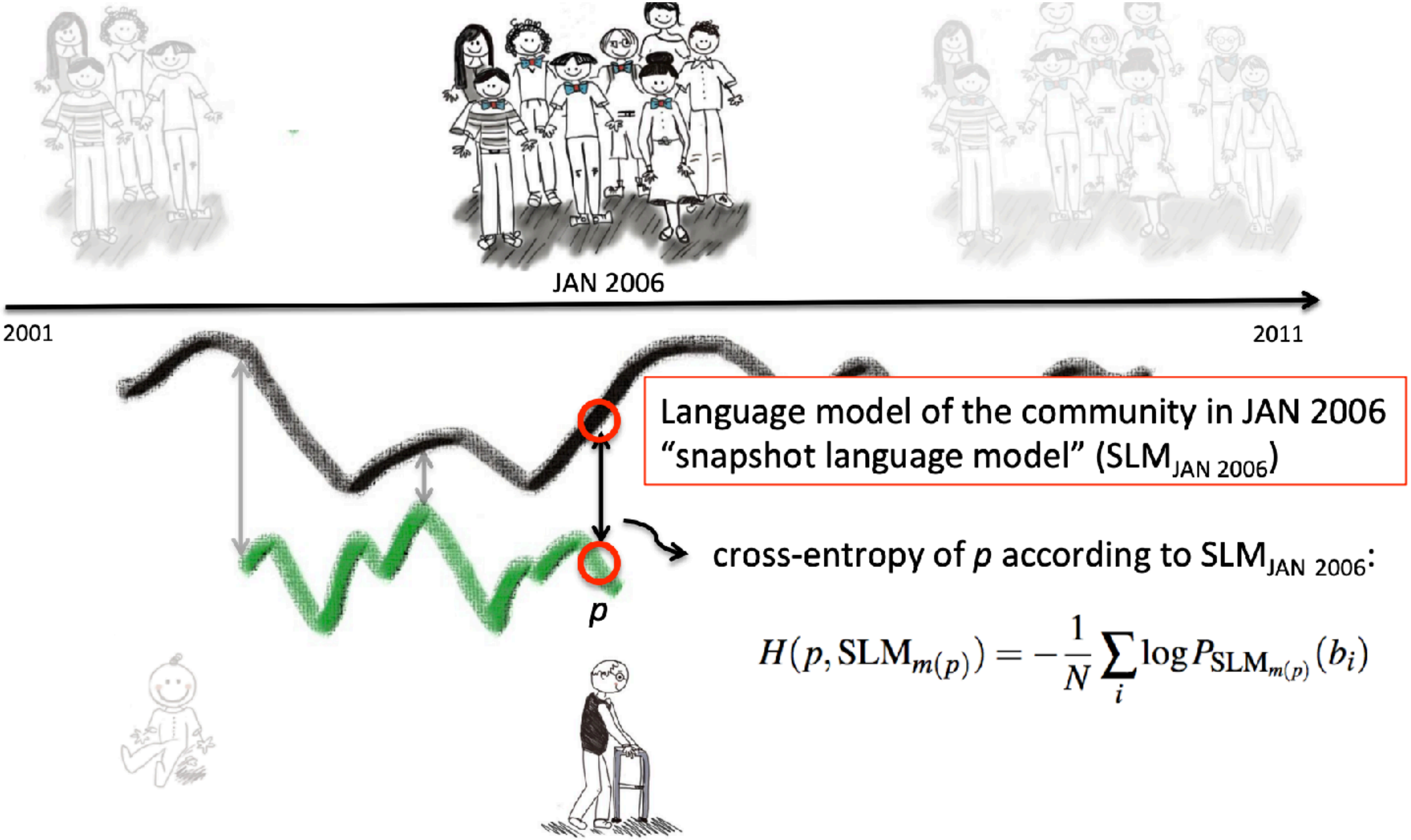


Example of user-level change:
Decrease in usage of 1st person pronouns
(e.g., I, me, mine, myself)

A sign of increasing identification with the community [Pennebaker 2007; Sherblom 2009]



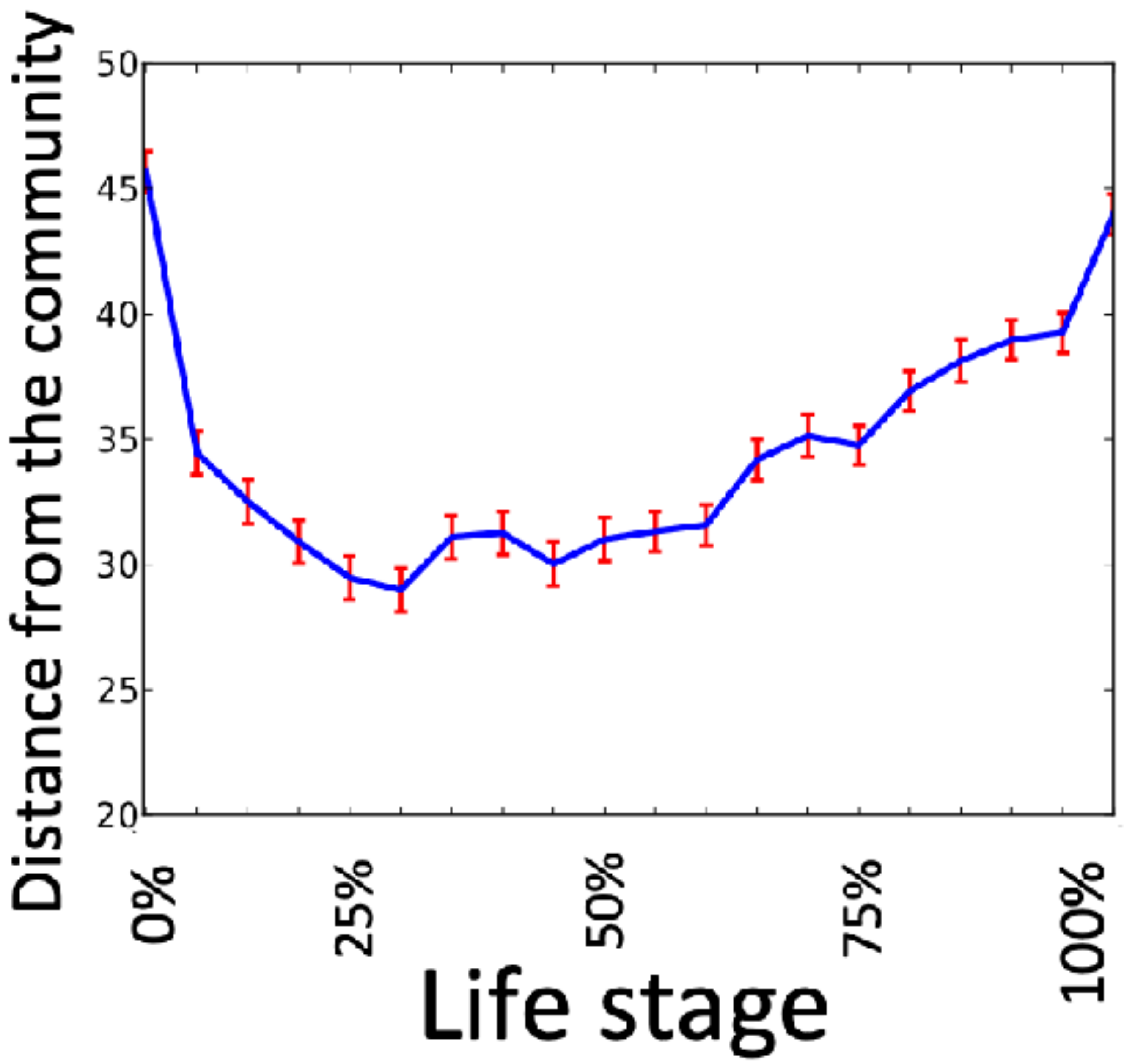
Distance from the community



$$H(p, \mathrm{SLM}_{m(p)}) = -\frac{1}{N}\sum_i \log P_{\mathrm{SLM}_{m(p)}}(b_i)$$

Distance from the community

Stage 1:
user **assimilates** the language of the community



Stage 2:
User's language **distances** itself from that of the community

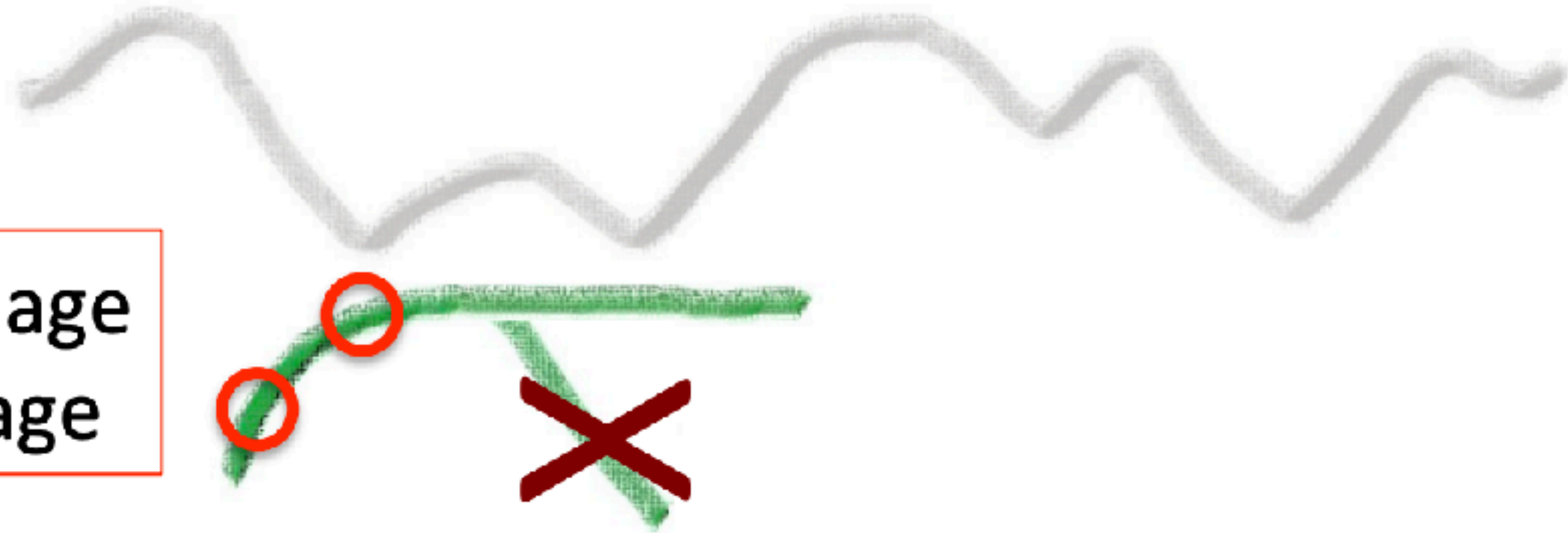
User-level stability

Compare user language
with her past language

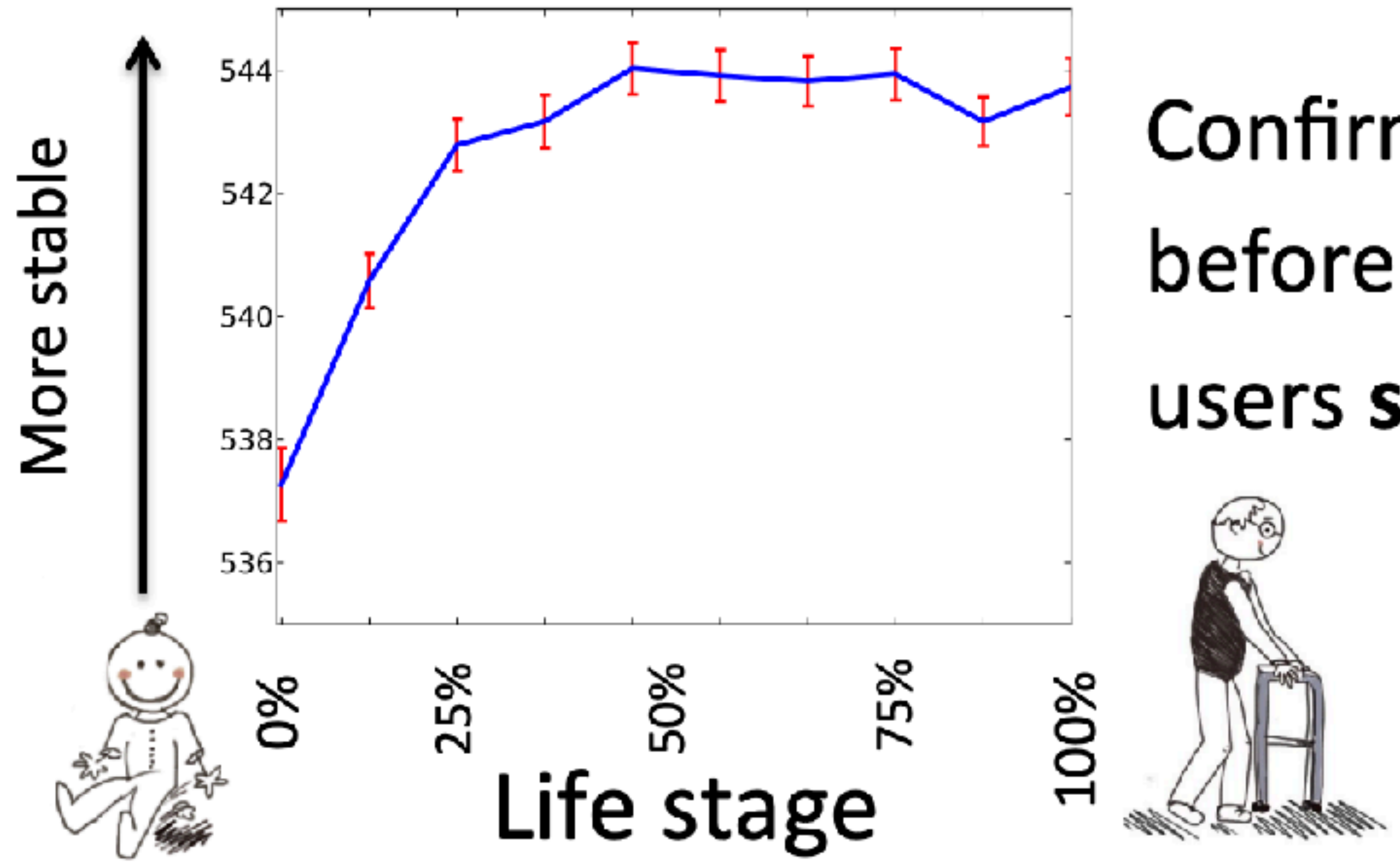
Hypothesis 1: User moves away from the community by **using innovative language**

Hypothesis 2: User **stops adapting** and gets out of tune with the changing community

User-level stability



Compare user language
with her past language



Confirms Hypothesis 2:
before abandoning,
users **stop adapting**

Whose language counts?

- Gururangan et al. 2022. Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection.
- Language model evaluations guide the selection of pretraining data
- Evaluation metrics encode, implicitly or explicitly, certain language ideologies

Whose language counts?

Model	Pretraining Data Sources	Citation
ELMo	1B Word benchmark	(Peters et al., 2018)
GPT-1	BookCorpus	(Radford et al., 2018)
GPT-2	WebText	(Radford et al., 2019)
BERT	BookCorpus + Wikipedia	(Devlin et al., 2019)
RoBERTa	BookCorpus + Wikipedia + CC-news + OpenWebText + Stories	(Liu et al., 2019)
XL-Net	BookCorpus + Wikipedia + Giga5 + ClueWeb 2012-B + Common Crawl	(Yang et al., 2019)
ALBERT	BERT, RoBERTa, and XL-net's data sources	(Lan et al., 2020)
T5	Common Crawl (filtered)	(Raffel et al., 2020)
XLM-R	Common Crawl (filtered)	(Conneau et al., 2020)
BART	BookCorpus + Wikipedia	(Lewis et al., 2020)
GPT-3	Wikipedia + Books + WebText (expanded) + Common Crawl (filtered)	(Brown et al., 2020)
ELECTRA	BookCorpus + Wikipedia + Giga5 + ClueWeb 2012-B + Common Crawl	(Clark et al., 2020)
Megatron-Turing NLG	The Pile + Common Crawl (filtered) + RealNews + Stories	(Kharya and Alvi, 2021)
Switch-C	Common Crawl (filtered)	(Fedus et al., 2021)
Gopher	MassiveWeb + Books + Common Crawl (filtered) + News + GitHub + Wikipedia	(Rae et al., 2021)

Table 5: Overview of recent language models and their training corpora. All studies tend to draw from the same core data sources: Wikipedia, Books, News, or filtered web dumps.

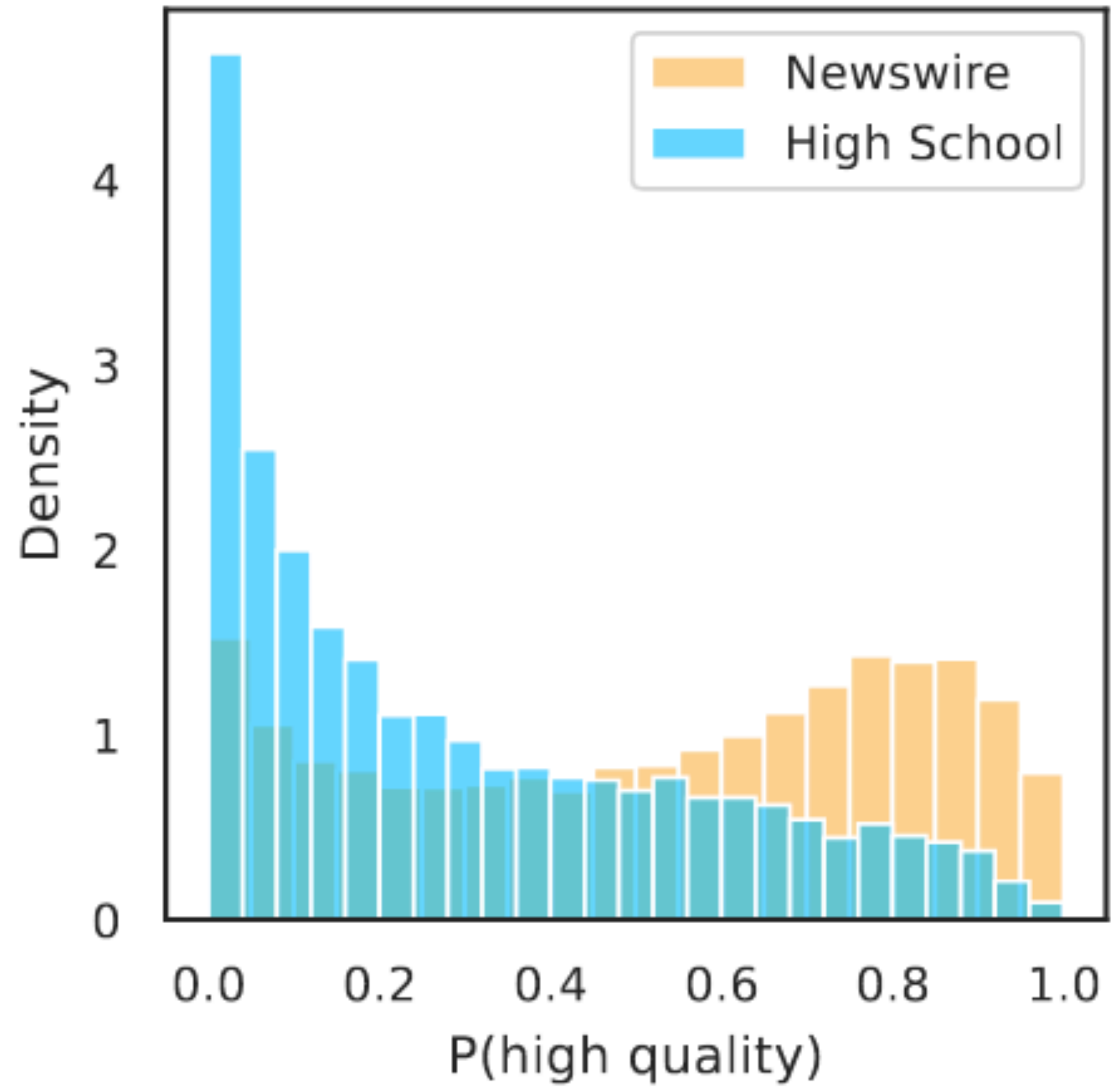
Whose language counts?

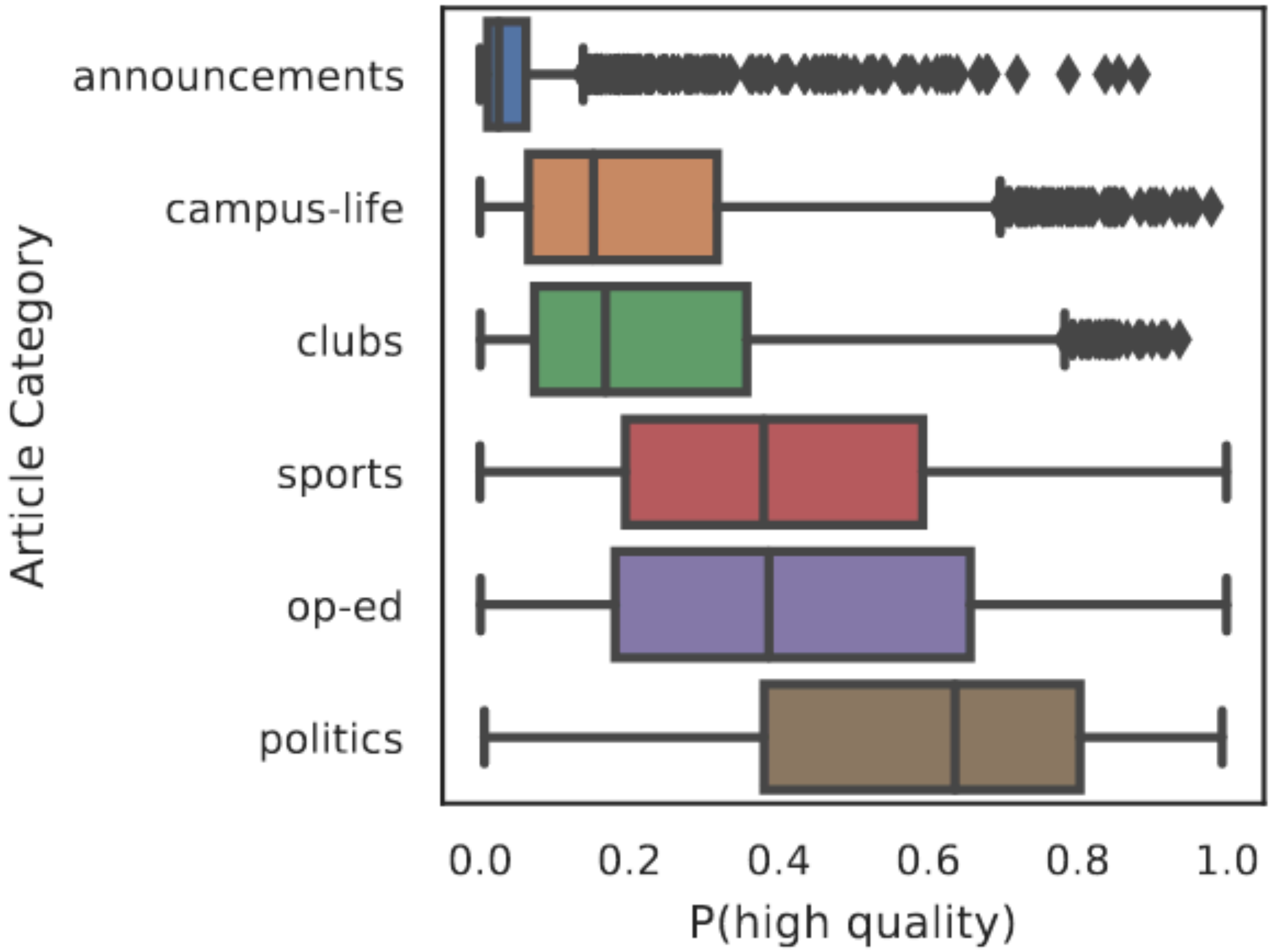
URL Domain	# Docs	% of Total Docs
bbc.co.uk	116K	1.50%
theguardian.com	115K	1.50%
washingtonpost.com	89K	1.20%
nytimes.com	88K	1.10%
reuters.com	79K	1.10%
huffingtonpost.com	72K	0.96%
cnn.com	70K	0.93%
cbc.ca	67K	0.89%
dailymail.co.uk	58K	0.77%
go.com	48K	0.63%

Table 1: The most popular top-level URL domains in OpenWebText. Mainstream news forms the overwhelming majority of content in the dataset. Overall, just 1% of the top-level URL domains in OpenWebText contribute 75% of the total documents in the corpus.

Whose language counts?

- Replicate GPT-3 quality filters
- Apply to diverse corpus of 2M high-school newspaper articles from after GPT-3 training
- Also apply to recent prizewinning books





Whose language counts?

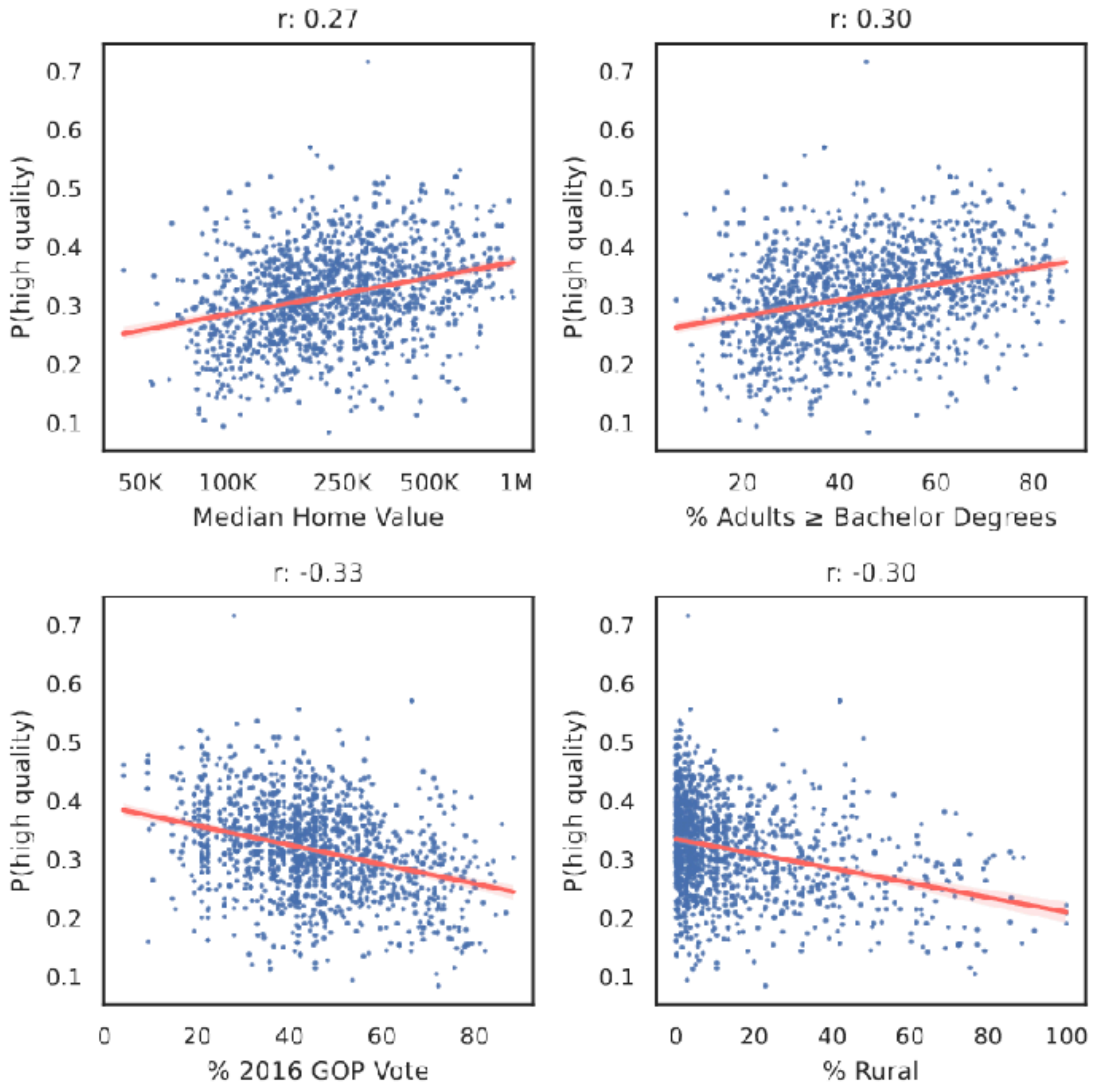


Figure 2: Scatter plots displaying correlations of select demographic features of a school's ZIP code or county with its average P(high quality).

Dependent variable: P(high quality)
Observations: 968 schools

Feature	Coefficient
Intercept	0.076
% Rural	-0.069^{***}
% Adults $\geq$ Bachelor Deg.	0.059^{**}
$\log_2$(Median Home Value)	0.010^{*}
$\log_2$(Number of students)	0.006^{*}
$\log_2$(Student:Teacher ratio)	-0.007
Is Public	0.015^{*}
Is Magnet	0.013
Is Charter	0.033
R^2	0.140
adj. R^2	0.133

Table 3: Regression of the average P(high quality) of a school in the U.S. SCHOOL NEWS dataset, on demographic variables. We observe that larger schools in educated, urban, and wealthy areas of the U.S tend to be scored higher by the GPT-3 quality filter. See §A.6 for more information on these features. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Whose language counts?

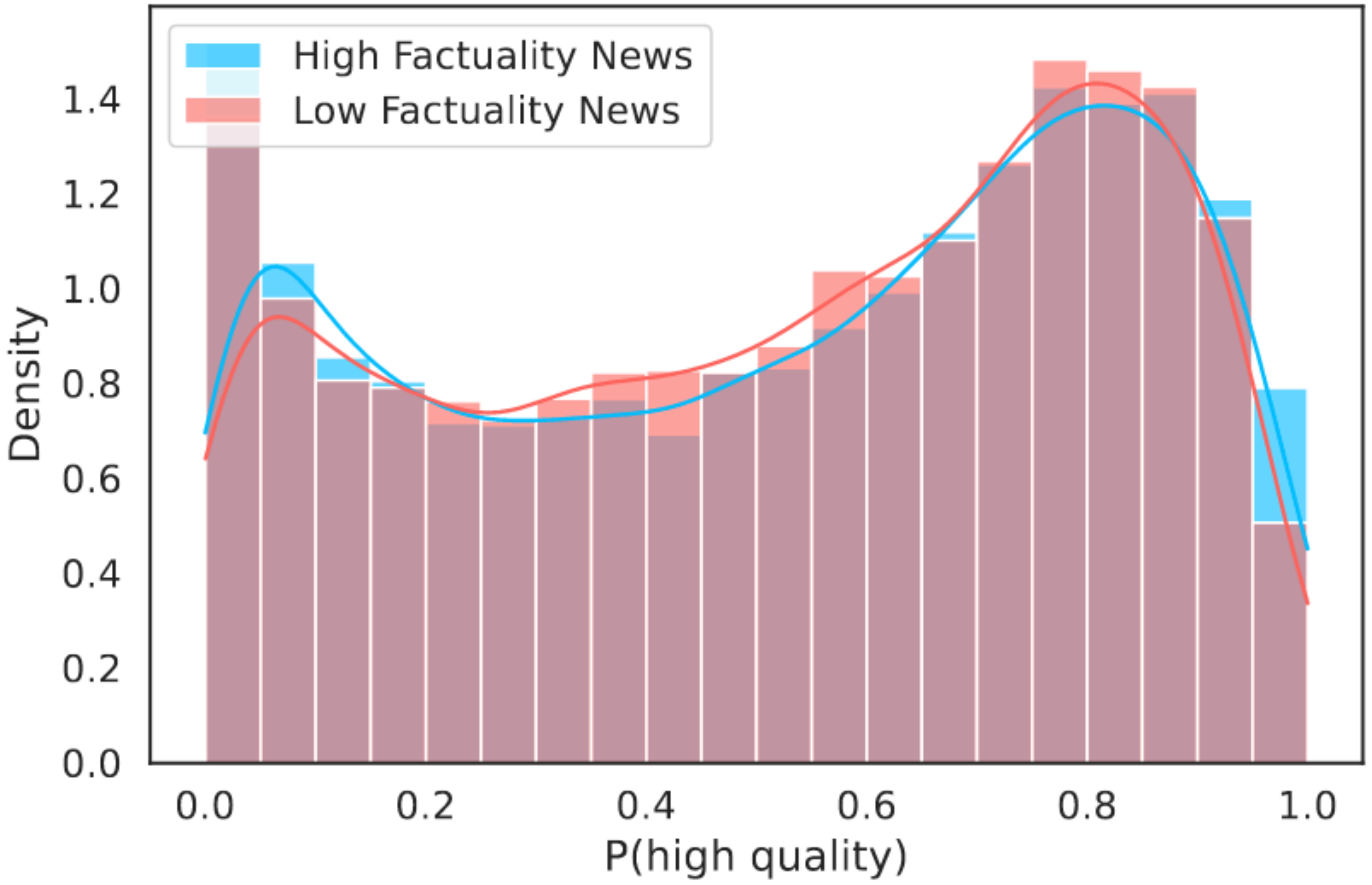


Figure 3: There is no difference in quality scores between articles written by news sources of high and low factual reliability.

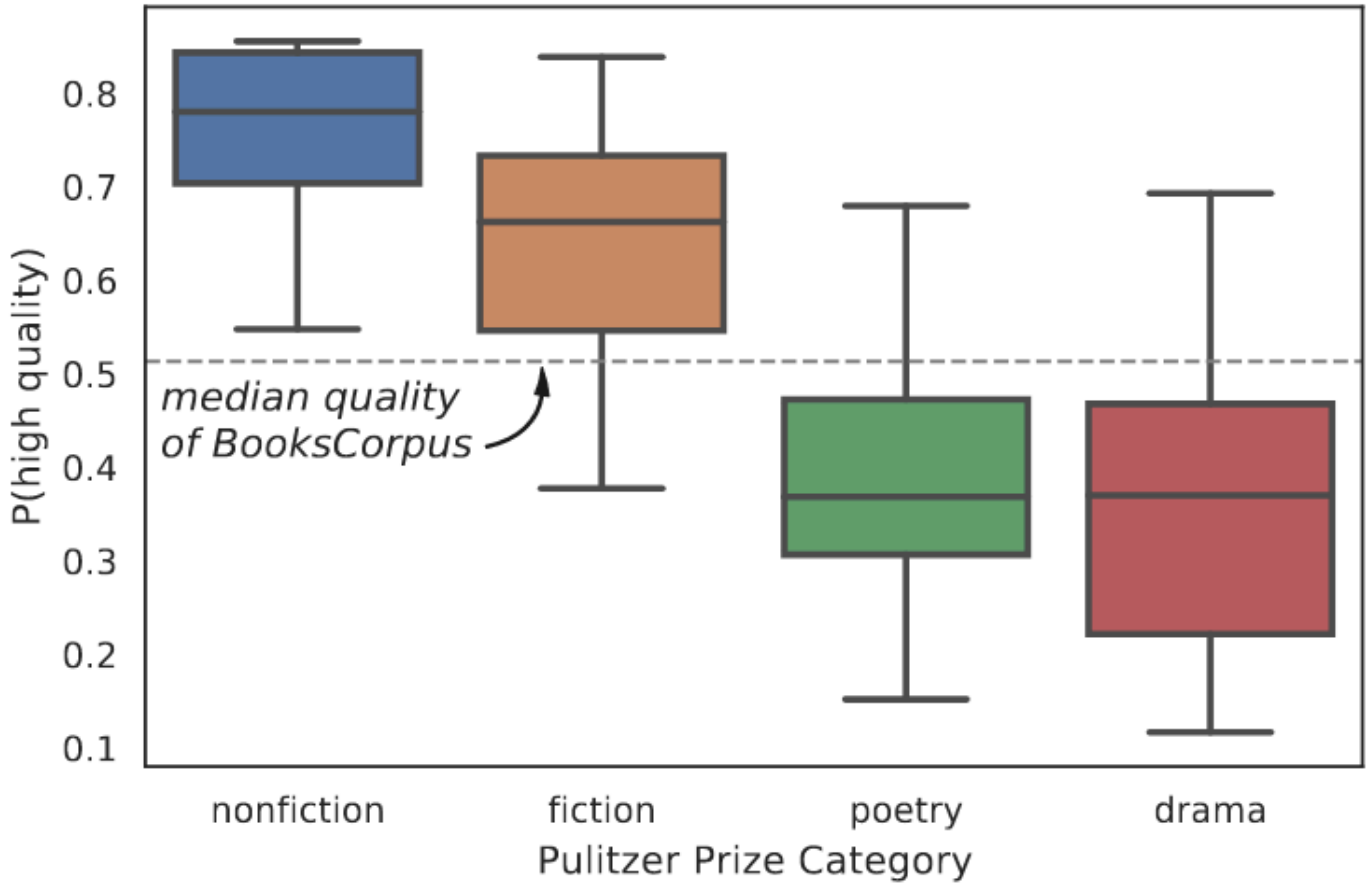


Figure 4: Among works that have won a Pulitzer Prize, the quality filter tends to favor nonfiction and longer fictional forms, disfavoring poetry and dramatic plays.

Whose language counts?

- Li et al. 2024. AboutMe: Using Self-Descriptions in Webpages to Document the Effects of English Pretraining Data Filters.
- Collect 10.3M author self-descriptions from websites
- Compare 10 different "quality" and language ID filters used by LLMs

Whose language counts?

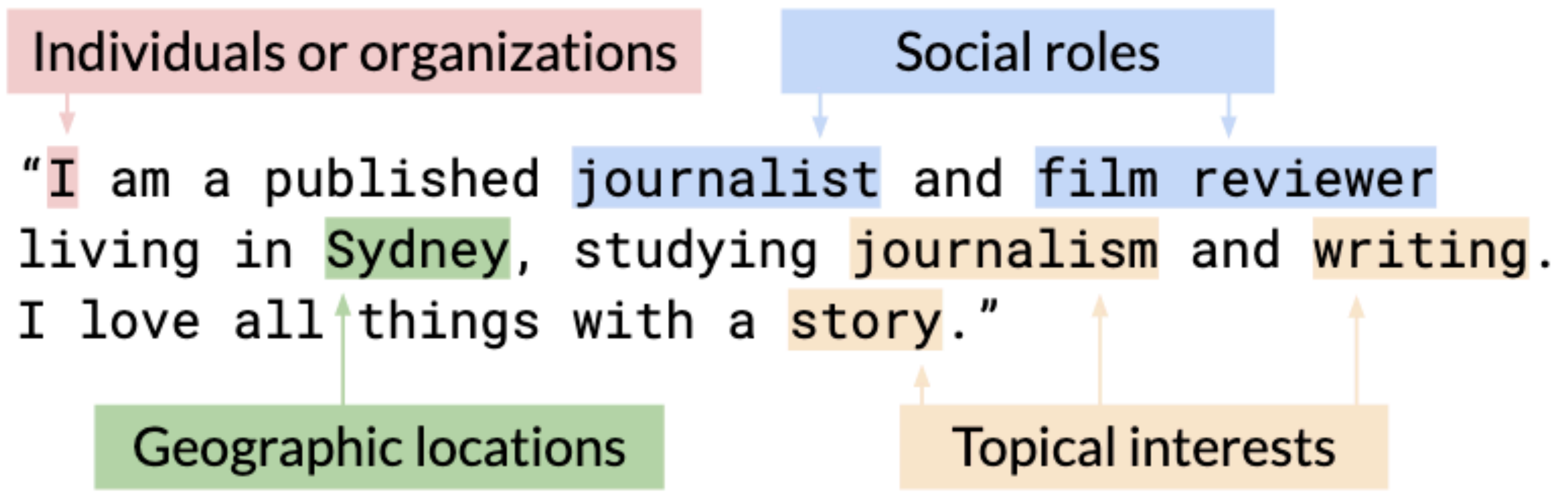


Figure 1: A paraphrased excerpt from a website's ABOUT page, with commonly stated social dimensions emphasized with highlighting.

Occupation family	Count	Examples of extracted roles
Arts, Design, Entertainment, Sports, & Media	1.1M	*artist*, *director*, *designer*, *writer*, *photographer*, *musician*, *player*
Production	620K	*designer*, *engineer*, *maker*, *builder*, *operator*, *mechanic*
Community & Social Service	452K	*therapist*, *educator*, *advisor*, *pastor*, *activist*, *social worker*
Computer & Mathematical	365K	*engineer*, *developer*, *scientist*, *strategist*, *programmer*
Educational Instruction & Library	308K	*teacher*, *professor*, *lecturer*, *curator*, *tutor*, *graduate student*

Table 2: Five most common occupation families in AboutMe, by website count, with example social roles. An extended version of this table is in Appendix F.3.

Whose language counts?

Filter	Examples of prior use	Removal strategy
⋆WIKIWEBBOOKS, or Wikipedia, OpenWebText, & Books3 classifier	GPT-3 (Brown et al., 2020)	Sampling based on scores
⋆OPENWEB, or Reddit outlinks classifier	the Pile (Gao et al., 2020)	Sampling based on scores
⋆WIKIREFS, or Wikipedia references classifier	LLaMA (Touvron et al., 2023a) & RedPajama (Computer, 2023)	Sampling based on scores
⋆WIKI, or Wikipedia classifier	Specified in reference mixes by Xie et al. (2023), PaLM (Chowdhery et al., 2023), and GPT-3 (Brown et al., 2020)	Sampling based on scores
⋆WIKI$_{ppl}$, or Wikipedia perplexity	CCNet (Wenzek et al., 2020)	Percentile cutoffs: 33.3% or 66.7%
⋆GOPHER length, wordlist, repetition, & symbol rules	Gopher (Rae et al., 2021), Chinchilla (Hoffmann et al., 2022), & RefinedWeb (Penedo et al., 2023)	Specific cutoffs for each rule
⋆**fastText** classifier	CCNet (Wenzek et al., 2020), LLaMA (Touvron et al., 2023a), RefinedWeb (Penedo et al., 2023)	Cutoffs: 0.50 (CCNet, LLaMA), 0.65 (RefinedWeb)
⋆**CLD2** classifier	The Pile (Gao et al., 2020)	Cutoff: 0.50
⋆**CLD3** classifier	multilingual C4 (Xue et al., 2021)	Cutoff: 0.70
⋆**langdetect** classifier	C4 (Dodge et al., 2021; Raffel et al., 2023)	Cutoff: 0.99

Whose language counts?

Topical interests				Social roles				Geography			
least	**– rate**	**most**	**– rate**	**least**	**– rate**	**most**	**– rate**	**least**	**– rate**	**most**	**– rate**
law, legal	0.19	fashion, women	0.47	counsellor	0.16	jewelry designer	0.42	Northern Europe	0.26	Eastern Asia	0.31
blog, like	0.19	furniture, jewelry	0.42	hypnotherapist	0.16	production designer	0.40	Central Asia	0.26	Southern Asia	0.30
insurance, care	0.20	online, store	0.40	atheist	0.16	retoucher	0.40	Western Europe	0.26	South-eastern Asia	0.29
financial, clients	0.20	com, www	0.39	executive coach	0.17	illustrator	0.38	Northern America	0.26	Northern Africa	0.29
solutions, technology	0.20	products, quality	0.37	psychotherapist	0.17	concept artist	0.38	Australia & NZ	0.27	Western Asia	0.29

Table 4: The topical clusters, social roles, and geographic subregions that are least and most filtered by GOPHER heuristics. Appendix B.2 describes how individual rules affect webpages.

Whose language counts?

Quality: WikiWebBooks				Quality: OpenWeb				Quality: WikiRefs			
↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate
news, media	0.27	home, homes	0.21	news, media	0.32	estate, real	0.20	news, media	0.28	blog, like	0.21
film, production	0.24	estate, real	0.18	writing, books	0.20	home, homes	0.18	club, members	0.23	furniture, jewelry	0.20
writing, books	0.24	service, cleaning	0.18	software, data	0.20	furniture, jewelry	0.17	music, band	0.23	home, homes	0.19
research, university	0.22	blog, like	0.16	like, love	0.18	fashion, women	0.17	film, production	0.23	fashion, women	0.19
music, band	0.21	insurance, care	0.16	site, information	0.18	blog, like	0.16	research, university	0.22	service, cleaning	0.18
Quality: Wiki				**Quality: Wiki$_{ppl}$**				**English: fastText**			
↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate
research, university	0.26	service, cleaning	0.22	law, legal	0.24	fashion, women	0.24	blog, like	0.22	fashion, women	0.21
film, production	0.25	home, homes	0.20	research, university	0.20	online, store	0.23	writing, books	0.22	online, store	0.20
music, band	0.21	insurance, care	0.16	god, church	0.19	quality, equipment	0.21	god, church	0.21	quality, equipment	0.18
art, gallery	0.21	marketing, digital	0.16	music, band	0.18	products, quality	0.21	photography, photographer	0.19	products, quality	0.18
law, legal	0.18	event, events	0.15	film, production	0.17	furniture, jewelry	0.20	like, love	0.19	furniture, jewelry	0.17
English: CLD2				**English: CLD3**				**English: langdetect**			
↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate
insurance, care	0.97	quality, equipment	0.13	service, cleaning	0.22	fashion, women	0.19	blog, like	0.94	online, store	0.11
service, cleaning	0.97	company, products	0.09	life, yoga	0.19	quality, equipment	0.17	writing, books	0.93	fashion, women	0.11
law, legal	0.97	energy, water	0.09	like, love	0.18	online, store	0.17	life, yoga	0.93	quality, equipment	0.11
financial, clients	0.97	com, www	0.09	blog, like	0.18	art, gallery	0.16	god, church	0.93	products, quality	0.11
home, homes	0.97	research, university	0.08	dog, pet	0.17	products, quality	0.15	law, legal	0.93	com, www	0.11

Table 5: The result of simulating two contrasting filtering scenarios: which topical interests are *most retained* when all pages except those with the highest scores are filtered (↑ *retained*), and which are *most removed* when pages with the lowest scores are filtered (↓ *removed*). Numeric columns are topics' page removal (–) or retained rate (+). A few topical interests that recur throughout the table are highlighted for clarity. See Appendix C.2 for an extended and more detailed version of this table.

Whose language counts?

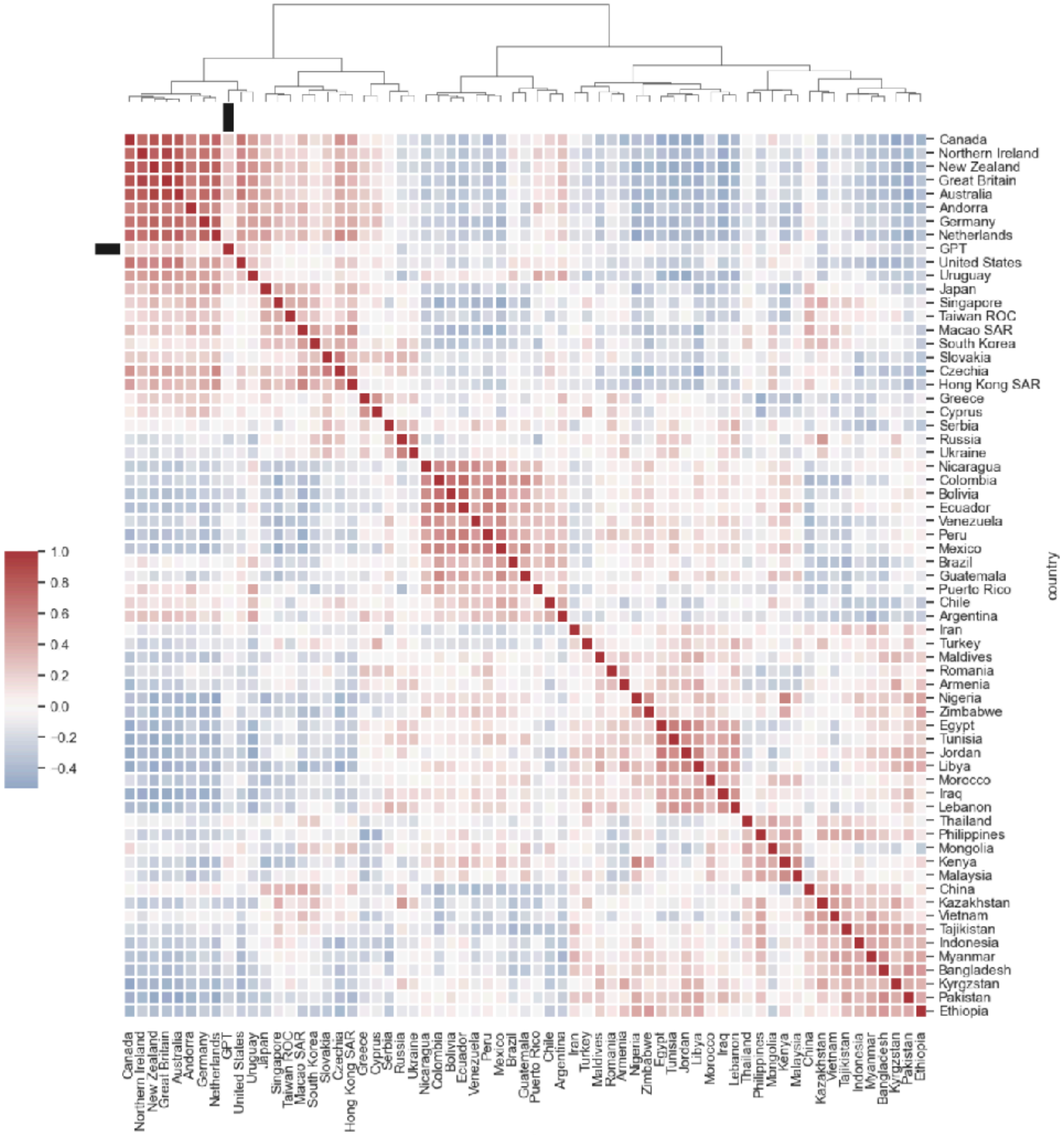
Quality: WIKIWEBBOOKS				Quality: OPENWEB				Quality: WIKIREFS			
↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate
correspondent	0.38	home inspector	0.33	game developer	0.43	home inspector	0.31	correspondent	0.32	quilter	0.25
game developer	0.37	realtor	0.24	game designer	0.39	residential specialist	0.27	mayor	0.30	home inspector	0.24
game designer	0.36	real estate agent	0.23	data scientist	0.35	realtor	0.26	co-writer	0.30	crafter	0.24
essayist	0.34	inspector	0.23	correspondent	0.32	real estate broker	0.25	historian	0.30	stager	0.22
historian	0.34	stager	0.21	software engineer	0.34	real estate agent	0.25	bandleader	0.30	jewelry designer	0.21
Quality: WIKI				**Quality: WIKI$_{ppl}$**				**English: fastText**			
↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate
laureate	0.35	wedding planner	0.21	law clerk	0.30	jewelry designer	0.17	christian	0.32	lighting designer	0.19
soprano	0.33	home inspector	0.20	litigator	0.26	lighting designer	0.16	catholic	0.31	production designer	0.18
conductor	0.32	momma	0.20	vice-chair	0.25	fashion designer	0.15	missionary	0.31	cinematographer	0.16
composer	0.31	dental assistant	0.20	conductor	0.24	production designer	0.14	mummy	0.29	retoucher	0.15
artistic director	0.30	mama	0.19	deputy	0.24	cinematographer	0.14	youth pastor	0.29	jewelry designer	0.15
English: CLD2				**English: CLD3**				**English: langdetect**			
↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate	↑ retained	+ rate	↓ removed	− rate
content strategist	0.99	laureate	0.13	counsellor	0.30	lighting designer	0.24	witch	0.96	production designer	0.11
home inspector	0.99	disciple	0.10	celebrant	0.28	production designer	0.23	barista	0.95	laureate	0.11
celebrant	0.99	soprano	0.10	hypnotherapist	0.25	sideman	0.21	naturopath	0.95	cinematographer	0.11
licensed professional counselor	0.98	language teacher	0.09	mummy	0.23	cinematographer	0.20	ally	0.95	retoucher	0.11
notary public	0.98	conductor	0.09	psychic	0.23	retoucher	0.19	cleaner	0.95	sideman	0.11

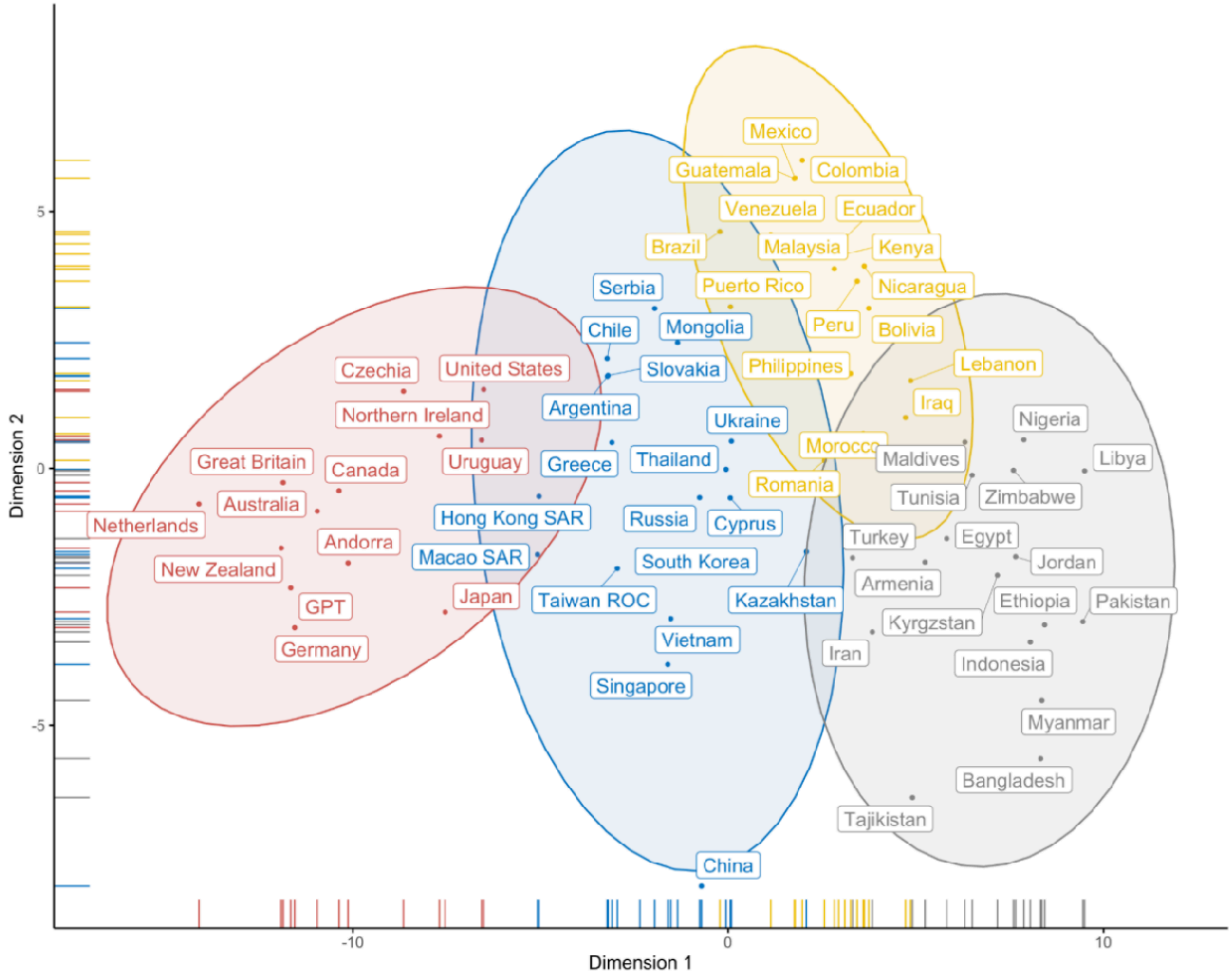
Occ. families: Arts, Design, Entertainment, Sports, & Media ; Community & Social Service ; Computer & Mathematical ; Sales & Related

Table 6: The result of simulating two contrasting filtering scenarios: which social roles are *most retained* when all pages except those with the highest scores are filtered (↑ *retained*), and which are *most removed* when pages with the lowest scores are filtered (↓ *removed*). Numeric columns include roles' page removal (−) or retained rate (+). For interpretation clarity, roles are highlighted if they belong to four frequently recurring O*NET occupation families. See Appendix F.4 for an extended and more detailed version of this table.

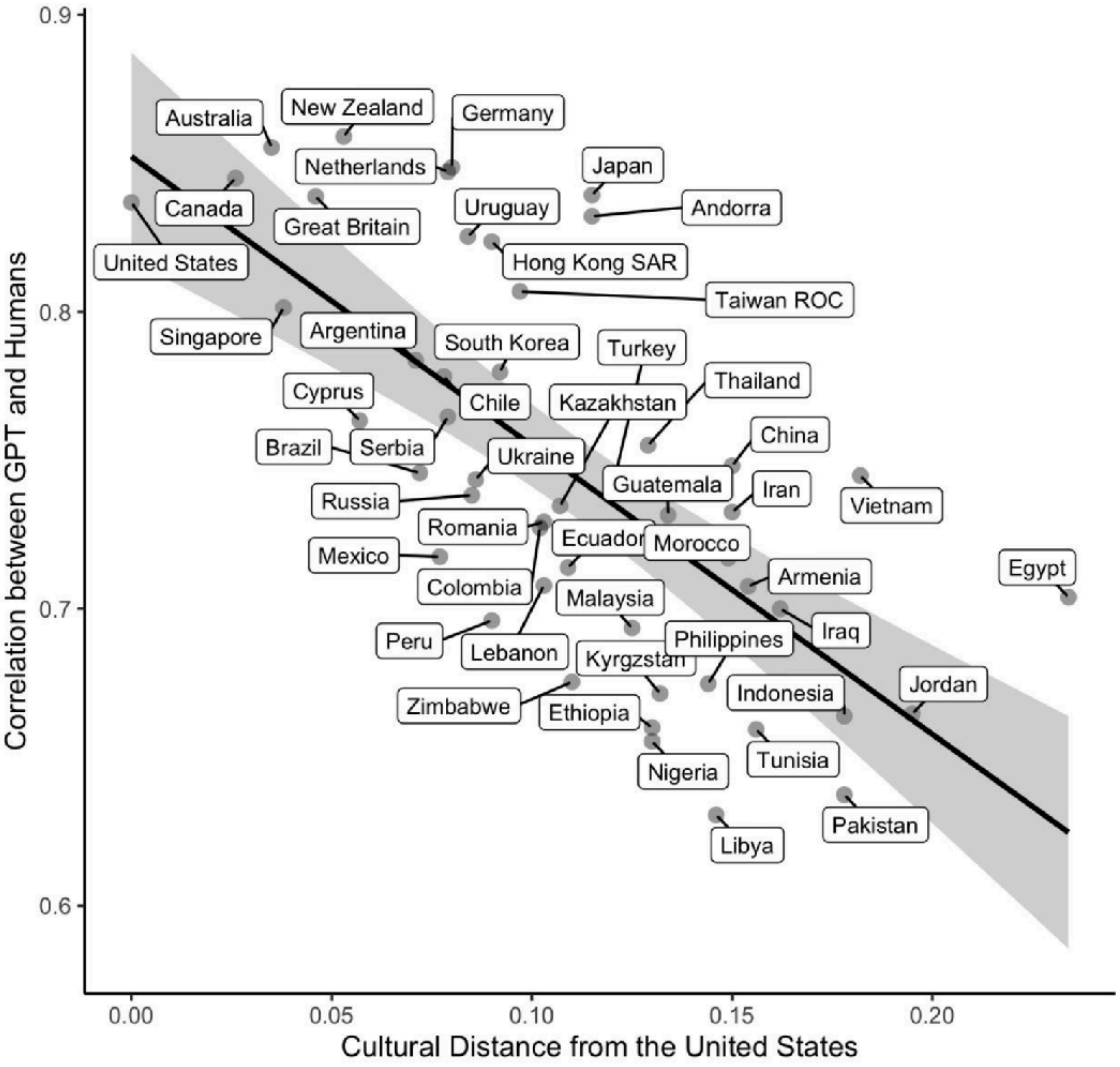
Populations of humans and machines

- Atari et al. 2023. Which humans?
- Many researchers use LLMs as proxies for human judgment
- LLM-based evaluation metrics rest on correlations between human and machine responses
- But human populations vary a lot! Some are WEIRD: Western, Educated, Industrialized, Rich, and Democratic
- Compare humans and models on World Values Survey (Haerpfer et al. 2020)


Canada
Northern Ireland
New Zealand
Great Britain
Australia
Andorra
Germany
Netherlands
GPT
United States
Uruguay
Japan
Singapore
Taiwan ROC
Macao SAR
South Korea
Slovakia
Czechia
Hong Kong SAR
Greece
Cyprus
Serbia
Russia
Ukraine
Nicaragua
Colombia
Bolivia
Ecuador
Venezuela
Peru
Mexico
Brazil
Guatemala
Puerto Rico
Chile
Argentina
Iran
Turkey
Maldives
Romania
Armenia
Nigeria
Zimbabwe
Egypt
Tunisia
Jordan
Libya
Morocco
Iraq
Lebanon
Thailand
Philippines
Mongolia
Kenya
Malaysia
China
Kazakhstan
Vietnam
Tajikistan
Indonesia
Myanmar
Bangladesh
Kyrgyzstan
Pakistan
Ethiopia
country
country
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4



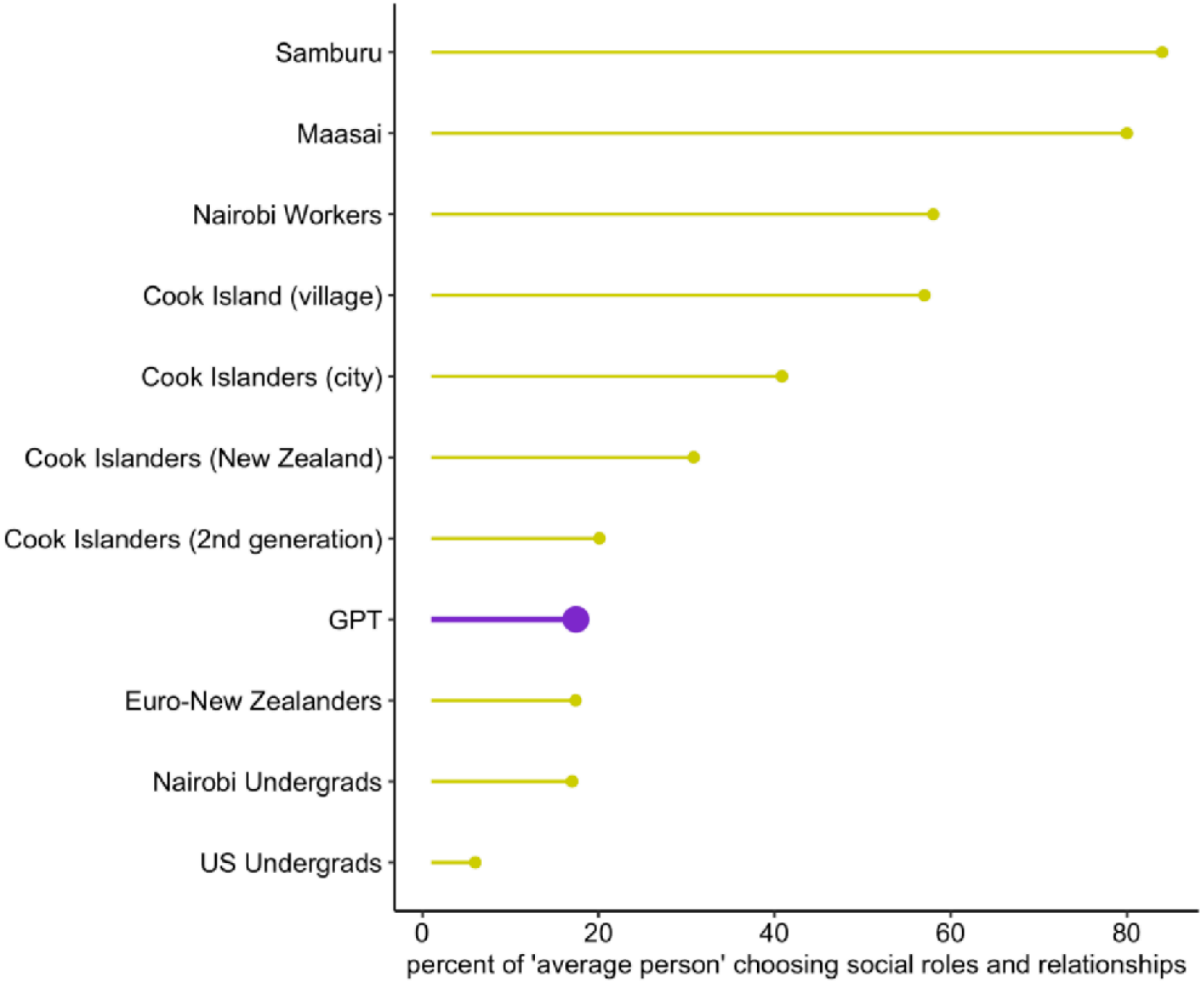
Mexico
Guatemala
Colombia
Venezuela
Ecuador
Brazil
Malaysia
Kenya
Serbia
Puerto Rico
Nicaragua
Chile
Mongolia
Peru
Bolivia
Czechia
United States
Slovakia
Philippines
Lebanon
Northern Ireland
Argentina
Iraq
Ukraine
Nigeria
Great Britain
Canada
Uruguay
Greece
Thailand
Morocco
Maldives
Libya
Australia
Romania
Tunisia
Zimbabwe
Netherlands
Hong Kong SAR
Russia
Cyprus
Andorra
Turkey
Egypt
New Zealand
Macao SAR
South Korea
Jordan
Armenia
GPT
Japan
Taiwan ROC
Kazakhstan
Ethiopia
Pakistan
Kyrgzstan
Germany
Vietnam
Iran
Indonesia
Singapore
Myanmar
Bangladesh
Tajikistan
China
Dimension 2
Dimension 1
5
0
-5
-10
0
10



Correlation between GPT and Humans
0.9
0.8
0.7
0.6
0.00
0.05
0.10
0.15
0.20
Cultural Distance from the United States
Australia
New Zealand
Germany
Netherlands
Japan
Canada
Andorra
Uruguay
Great Britain
United States
Hong Kong SAR
Taiwan ROC
Singapore
Argentina
South Korea
Turkey
Thailand
Cyprus
Chile
Kazakhstan
China
Brazil
Serbia
Ukraine
Guatemala
Iran
Vietnam
Russia
Romania
Ecuador
Morocco
Mexico
Egypt
Armenia
Colombia
Malaysia
Iraq
Peru
Lebanon
Philippines
Kyrgzstan
Jordan
Indonesia
Zimbabwe
Ethiopia
Tunisia
Nigeria
Pakistan
Libya

Populations of humans and machines

- "List 10 specific ways that an average person may choose to identify themselves. Start with 'I am…'"



Summary

- Language is a technology with a functional role in cultural transmission and coordination
- Language reflects human identities, relationships, and power
- Computational models can help us map these social phenomena
- Language technologies reproduce social phenomena