THE LAWS OF "LOL": COMPUTATIONAL APPROACHES TO SOCIOLINGUISTIC VARIATION IN ONLINE DISCUSSIONS

A Dissertation
Presented to
The Academic Faculty

By

Ian Stewart

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Computer Science
Interactive Computing

Georgia Institute of Technology

December 2020

THE LAWS OF "LOL": COMPUTATIONAL APPROACHES TO SOCIOLINGUISTIC VARIATION IN ONLINE DISCUSSIONS

Thesis committee:

Jacob Eisenstein, Advisor
School of Interactive Computing
Georgia Institute of Technology

Mark Riedl
School of Interactive Computing
Georgia Institute of Technology

Diyi Yang, Advisor
School of Interactive Computing
Georgia Institute of Technology

David Jurgens
School of Information
University of Michigan

Munmun De Choudhury
School of Interactive Computing
Georgia Institute of Technology

Timothy Baldwin
School of Computing and Information Systems
University of Melbourne

Date approved: 10 August 2020

Like the rest of language, variation does not simply reflect the social, but enacts it, and in the course of this enactment, it participates in social change.

Penelope Eckert

To all the kids who asked: “But what does that word *really* mean?” - Stay curious.

ACKNOWLEDGMENTS

Getting paid to do research in sociolinguistics is a remarkable privilege. The field has lost too many bright young researchers with valuable perspectives due to lack of support. No matter where life leads me, I am grateful to have had this opportunity.

That being said, doing research in computational sociolinguistics has not been easy. At the time of writing, despite repeated contributions from brilliant researchers, it is still not a very well-accepted subfield within computational linguistics. As I approach the end of the PhD, I realize how many people have gone out of their way to help me in this research area over the years. While I didn't have the "join a big lab, collaborate, profit" experience that is the norm in computer science, I have been lucky to build deep relationships with a few people who pushed me to become a better researcher. If I can pay forward even half of what I have been given, then I will consider myself a lucky man.

To my advisor Jacob Eisenstein: thank you for everything. Even before I came to Georgia Tech as a student, you were in my corner. You stuck by my research when the outcome wasn't certain, you helped me change directions when I needed to change, and you made me think hard about what "sociolinguistics" really means. I promise that I won't ever leave out a control condition again! You supported me in times when I really needed it, especially in the final few years of the PhD. If I get to mentor students in the future, I hope that I can pass on the rigor and creativity that you impressed onto me.

To my advisor Diyi Yang: thank you for stepping into an uncertain situation and putting in the work to get me across the finish line. You pushed me to connect my research to bigger problems in social computing, and you absorbed a lot of information quickly to get up to speed with my work. I think that you've cumulatively read about 300 pages of text from me, which is 300 more than I should have asked for. It was a joy to be a small part of the SALT lab, and I can't wait to see how it grows in the future.

To the other committee members, Munmun De Choudhury, Mark Riedl, David

Jurgens, and Tim Baldwin: thank you for all your feedback on my work. This thesis has come a long way thanks to your help and your advice on how to see the forest for the trees, from the early stages of the proposal to the final document here. I was lucky to have a diverse cast of researchers to provide a variety of different insights into this research, from methods to theory to applications. I will carry your advice with me into my next phase of life.

I am grateful for other faculty, senior researchers, and administrators who stepped in to help me over the years. For help on the loanword project: Cecilia Montes-Alcalá and Chad Howe, you have my gratitude as a non-Spanish expert. For advising me during internships: Svitlana Volkova and Emilio Zagheni, it was an honor to learn from you and blaze new trails in computational social science together. For letting me give guest lectures and helping me stay connected to linguistics: Lelia Glass, thanks for being the "linguist-in-residence" for all of us at GT. For interim advising: Amy Bruckman, thanks for supporting me in a time of uncertainty and holding down the social computing seminar over the years. For general support: Rosa Arriaga, you backed all us students without wavering, no matter how complicated our academic situations became. For all the help with reimbursements, research funding, and fulfilling program requirements: Renee Jamieson, Carolyn Daley-Foster, and Kyla Hanson (among others), you made the administrative part of PhD life easy. For organizing the computational sociolinguistics workshop in 2018: Dirk Hovy and Dong Nguyen, that workshop lit a fire in me that propelled me through the last leg of my PhD. For introducing me to academic research: Jim Stanford and Sravana Reddy, thanks for giving me the support to combine computer science and sociolinguistics as a scrappy undergrad. Lastly, for teaching me French, Latin, and so much more: Dr. Ladd, thanks for jump-starting my entry to the world of linguistics.

To my lab-mates: it's been real, and it's been fun. We may not have been the most traditional or tight-knit lab, but you all helped me learn about NLP, how to argue well as an academic, and how to do research that matters. I still owe you for the many practice

talks that I put you through! To Yangfeng, Yi, Han, Jiaao, and others, thanks for making the lab an energetic and cutting-edge environment for research. Umashanthi: you've been a research inspiration throughout my PhD, and I was lucky to absorb your insight on computational sociolinguistics and academic careers in general. To Yuval, thanks for the work on the Catalan paper, for all those lunches and coffee breaks, and for the marathon running/punning sessions. To Sarah, thanks for being a source of positive energy, spilling the tea when I needed news, and listening when I needed to talk. To Sandeep, I can't thank you enough. From day 1, we had rollicking conversations about sociolinguistics and social science in general. We've gotten to talk over each other in reading groups, go to conferences, tag-team a tutorial (on language change!), and test-drive some really out-there research ideas. You made me think hard about my own interests and told me what I needed to hear, even if it wasn't what I wanted to hear at the time. I've been changed for the better. Best of luck to you all.

To my fellow PhD travelers, it's been a privilege learning with you all. Despite knowing nothing about social computing before I started, you gave me a new perspective on deep issues concerning communities, social norms, and mental health. To Sindhu, Koustuv, and Eshwar, I had a blast talking about research and life with you, and I hope our paths keep crossing at conferences. To Stevie, thanks for your help with the hashtag variation paper, your candid career advice, and all the fun parties. To Sid and Kelsey, thanks for being the best roommates I could've asked for. I'll miss eating, working, laughing, and learning with you both. To Dave and Lara, thanks for all the fun game nights - I promise I'll be DM sometime. To Ari, thanks for all the goofy movie nights, dinner dates, and general tomfoolery. To Emily and Fred, thanks for the many Fremilunchians and venting sessions (I still don't have a Which Wich account). To Su Lin, you helped me stay connected to the core sociolinguistics research program that drives this thesis, which for someone in a social computing department was more important than you may know. To the other students in the Human-Centered Computing program and in computer

science at GT: I'm glad that our paths crossed, and I hope you stay brilliant and creative.

I don't know what the equivalent of the "red pill" is among PhD students, but it definitely exists. The friends outside of academia kept me grounded and helped stop me from going too far down the rabbit hole. To Meaghan, Marissa, Ryan, Xavier, Cami, and Alex, thanks for the monthly video chats and too-rare reunions in various cities. To Abby and Caleb, thanks for visiting me in Atlanta and encouraging me to invest in myself outside of academic pursuits. To the long list of other friends outside the ivory tower, thanks for listening to my stories about research: I owe you a drink.

Since I first started learning other languages, my family has enabled my linguistic obsessions with both good and bad results. For me and my family, a scientist is not a rare or special category of human; a scientist just needs to be curious about the world and to direct that curiosity into observing, analyzing, and explaining interesting patterns in the world. I know that linguistics isn't the "normal" kind of science that you can put under a microscope, but it's the right kind for me. Andrea: you've always been a role model for me as someone who puts science into action and stands up for what you believe in. Lily: you are the light of our family and you give me such hope, even when you give me grief. Mom and Dad: you always made sure that learning comes first and that everything else follows. You gave us the love and support that we needed to become our best selves, no matter how weird our interests may have seemed. To my family: this one's for you.

The work in this thesis has been funded by grants through the Air Force Research Laboratory, the National Science Foundation, and the National Institutes of Health. The opinions expressed in this document do not reflect the opinions of the funding organizations. I thank all the reviewers, editors and annotators who helped shape the work in this thesis. All remaining errors are mine alone. I thank all the participants who provided the social media data for this thesis, and I hope that this work can help make social media platforms (slightly) better.

SUMMARY

When speaking or writing, a person often chooses one form of language over another based on social constraints, including expectations in a conversation, participation in a global change, or expression of underlying attitudes. Sociolinguistic variation (e.g. choosing *going* versus *goin'*) can reveal consistent social differences such as dialects and consistent social motivations such as audience design. While traditional sociolinguistics studies variation in spoken communication, computational sociolinguistics investigates written communication on social media. The structured nature of online discussions and the diversity of language patterns allow computational sociolinguists to test highly specific hypotheses about communication, such different configurations of listener "audience." Studying communication choices in online discussions sheds light on long-standing sociolinguistic questions that are hard to tackle, and helps social media platforms anticipate their members' complicated patterns of participation in conversations.

To that end, this thesis explores open questions in sociolinguistic research by quantifying language variation patterns in online discussions. I leverage the "birds-eye" view of social media to focus on three major questions in sociolinguistics research relating to authors' participation in online discussions. First, I test the role of conversation expectations in the context of content bans and crisis events, and I show that authors vary their language to adjust to audience expectations in line with community standards and shared knowledge. Next, I investigate language change in online discussions and show that language structure, more than social context, explains word adoption. Lastly, I investigate the expression of social attitudes among multilingual speakers, and I find that such attitudes can explain language choice when the attitudes have a clear social meaning based on the discussion context. This thesis demonstrates the rich opportunities that social media provides for addressing sociolinguistic questions and provides insight into how people adapt to the communication affordances in online platforms.

TABLE OF CONTENTS

Acknowledgments v

Summary ix

List of Tables xvi

List of Figures xix

Chapter 1: Introduction 1

1.1 Thesis organization 4

1.1.1 Research question motivations 4

1.1.2 Thesis studies 7

1.2 Thesis statement 12

1.3 Contributions of thesis 12

Chapter 2: Background 16

2.1 Language variation and change 16

2.1.1 Variation in structure 17

2.1.2 Sociolinguistic theory: speaker-level motivations 19

2.1.3 Sociolinguistic theory: language change 24

2.2 Language variation on the internet 26

2.2.1 Computational sociolinguistics 27

2.2.2 Social computing 28

Chapter 3: Adoption of variant hashtags in online communities 32

3.1 Motivation 33

3.2 Data 35

3.2.1 Data collection 35

3.2.2 Feature extraction 37

3.3 Methods 38

3.3.1 Definitions 38

3.3.2 Measuring orthographic variation: edit distance 40

3.3.3 Statistical models 40

3.4 Results 42

3.4.1 Orthographic variant authors 43

3.4.2 Differences in variant depth by membership 45

3.4.3 Social reception of different variants 47

3.5 Limitations and future work 48

3.6 Contributions 50

Chapter 4: Characterizing collective attention through descriptor phrases . . . 52

4.1 Motivation 53

4.2 Data 56

4.2.1 Collection 57

4.2.2 Extracting and filtering locations 58

4.2.3 Extracting descriptor phrases . . . 59
4.3 Results . . . 62
4.3.1 Non-temporal social factors in descriptor use . . . 62
4.3.2 Collective change in descriptor context use . . . 65
4.3.3 Individual change in descriptor context use . . . 66
4.4 Limitations and future work . . . 68
4.5 Contributions . . . 70
4.6 Thesis section summary . . . 72

Chapter 5: Word growth and decline . . . 73

5.1 Motivation . . . 74
5.2 Data . . . 75
5.2.1 Finding growth words . . . 77
5.2.2 Finding decline words . . . 77
5.3 Methods . . . 79
5.3.1 Social dissemination . . . 80
5.3.2 Linguistic dissemination . . . 80
5.4 Results . . . 82
5.4.1 Correlational analysis . . . 82
5.4.2 Causal analysis . . . 83
5.4.3 Predictive analysis . . . 85
5.4.4 Survival analysis . . . 88
5.5 Limitations and future work . . . 90

5.6 Contributions 91

5.7 Thesis section summary 92

Chapter 6: Language choice in discussion of a political referendum 94

6.1 Motivation 95

6.2 Data 96

6.3 Results 98

6.3.1 Catalan usage and political attitude 98

6.3.2 Catalan usage, topic, and audience 100

6.4 Limitations and future work 103

6.5 Contributions 104

Chapter 7: Morphological integration of English loanwords in Spanish 106

7.1 Motivation 107

7.2 Data 112

7.2.1 Identifying loanwords 112

7.2.2 Identifying native verbs 113

7.2.3 Collecting loanword data 114

7.2.4 Identifying cultural media 115

7.2.5 Addressing confounds in media sharing 116

7.2.6 Data variable summary 121

7.3 Results 123

7.3.1 Differences in integration by domain 123

7.3.2 The role of demographics and behavior in integration 126

7.3.3 The role of media consumption in integration 129
7.4 Limitations and future work 132
7.5 Contributions 133
7.6 Thesis section summary 135

Chapter 8: Conclusion 136
8.1 Thesis summary 136
8.1.1 RQ1 results summary 136
8.1.2 RQ2 results summary 140
8.1.3 RQ3 results summary 141
8.1.4 Overall summary 145
8.1.5 Study design considerations 148
8.2 Limitations 152
8.3 Implications for future work 156
8.3.1 Practical applications 157
8.3.2 Theoretical considerations 159

Appendices 162

Appendix A: Collective attention 163
A.1 Detecting author social status 163
A.2 Robustness check for descriptor extraction 164

Appendix B: Loanword integration 165
B.1 All integrated and light verb pairs 165

References 169

LIST OF TABLES

1.1 Summary of the high-level research questions addressed by the thesis and the corresponding results. 11

3.1 Summary of orthographic variants grouped by edit distance. The edit distance 1 group has the greatest variety of source hashtags and unique variants, while the edit distance 4 group has the lowest variety. This study is restricted to variant hashtags with edit distance at or below 4, due to data sparsity above edit distance 4. 36

3.2 Regression results for variant appearance in a post, as predicted by relative time variables. *** indicates $p < 0.0001$. In all tables, β indicates the regression coefficient and S.E. indicates the standard error. 43

3.3 Regression results for variant appearance in a post, as predicted by the length of a member's lifespan (observed activity period). *** indicates $p < 0.0001$. 43

3.4 Logistic regression to predict the appearance of a variant with a specified edit distance, as predicted by (1) age and (2) lifespan. *** indicates $p < 0.0001$, * indicates $p < 0.05$. 47

3.5 Poisson regressions for social reception, as predicted by membership and language variables (hashtag coefficients omitted for brevity). *** indicates $p < 0.0001$, otherwise $p > 0.05$. Both models achieve a weak fit: the LOGCOMMENTS regression has R^2=6.82E-3 ($F = 107, p < 0.001$) and the LOGLIKES has R^2=0.0902 ($F = 1550, p < 0.001$). 48

4.1 Summary statistics for Twitter data. 57

4.2 Summary statistics for active authors on Twitter. 58

4.3 Phrase patterns to capture descriptor phrases in location mentions. Head location marked with underline, context location marked with double underline. 60

4.4 Summary of explanatory variables and corresponding metrics, used for descriptor phrase prediction. 61

4.5 Logistic regression results for all analysis, predicting the presence of a descriptor phrase. * indicates $p < 0.05$, otherwise $p > 0.05$. 64

5.1 Data summary statistics. 76

5.2 Examples of nonstandard words in all word sets: growth ($\mathcal{G}$), logistic decline ($\mathcal{D}_l$) and piecewise decline ($\mathcal{D}_p$). 79

5.3 Word formation category counts in growth ($\mathcal{G}$) and decline ($\mathcal{D}$) word sets. . 79

5.4 Percent of variance explained in frequency change, computed over all growth words $\mathcal{G}$. $N = 26,880$ for $k = 12$, $N = 13,440$ for $k = 24$. 83

5.5 Cox regression results for predicting word death with all predictors (`f+L+S`) averaged over first $k = 3$ months. *** indicates $p < 0.001$, otherwise $p > 0.05$. 89

6.1 Hashtags related to the Catalonian referendum, their attitudes (neutral/pro/anti) and their frequencies in the CT dataset. 97

6.2 Tweet and author counts for the attitude study. 99

6.3 Results of the attitude study. $d = \hat{p}_{pro} - \hat{p}_{anti}$. 100

6.4 Tweet and author counts for each condition in the topic/audience study. 'hash' stands for 'tweets with hashtags'. 101

6.5 Results of the topic/audience study. $\bar{d}_{\mathcal{U}_R}$ is the difference in rate of Catalan use between treatment settings and control settings, averaged across authors. 101

7.1 Top 5 most frequent loanwords and corresponding verb forms. 113

7.2 Top 5 most frequent native word pairs and corresponding verb forms. 114

7.3 Summary of all author-level variables used in study. 123

7.4 Loanwords with highest rate of integration difference between newspaper and social media writing. 126

7.5 Regression results for loanword and native word integrated verb prediction. *** indicates $p < 0.001$, ** indicates $p < 0.01$, $\sim$ indicates $p > 0.05$. . . . 128

7.6 Regression results for loanword and native word light verb prediction, with media variable. *** indicates $p < 0.001$, ** indicates $p < 0.01$, $\sim$ indicates $p > 0.05$. 130

A.1 Regression results for Facebook data in RQ1, using `spacy` parses to detect descriptor phrases. * indicates $p < 0.05$, otherwise $p > 0.05$. 164

LIST OF FIGURES

3.1 Summary histograms for all variables of interest, including relative time (e.g. DATE_RANGE), linguistic (MAX_EDIT) and social variables (LOGCOMMENTS). 35

3.2 Example timeline of member posts at times t_0 (first), t_i and t_n (final) that shows age with statistics SINCE_START and TILL_END, and showing lifespan with DATE_RANGE. 38

3.3 Frequency of variants over time, grouped by edit distance: e.g., DIST_1 tracks the normalized frequency of all posts with at least one variant with edit distance 1, such as *#anorexiaa*. 41

3.4 Probability of using a variant versus a member's age (weeks since first pro-ED post). 44

3.5 Distribution of maximum edit distances across all posts of specified member group (regular versus newcomer) at one week and 10 weeks after the ban (including average edit distance for each group). The newcomers used orthographic variants with consistently higher edit distances than the regulars. 45

3.6 Average edit distance over time, binned by DATE and DATE_RANGE and including 95% confidence intervals. 46

4.1 Example of collective attention expressed toward location mentions in discussion of various hurricanes on Twitter. Left y-axis (black solid line) indicates the location's log frequency, right y-axis (red dotted line) indicates the location's probability of receiving a descriptor phrase such as *San Juan, Puerto Rico*. For example, a 25% probability for "*San Juan*" means that 25% of all mentions of *San Juan* had a descriptor phrase. 54

5.1 Distribution of mean linguistic dissemination (D^L) across part of speech groups. 81

5.2 Average dose response function for all treatment variables, where outcome is probability of word growth. 95% confidence intervals plotted in red, chance rate of 50% marked with dotted black line. 85

5.3 Prediction accuracy for different feature sets using $k = 1...12$ months of training data. f indicates frequency-only, $f + L$ frequency plus linguistic dissemination, $f + S$ frequency plus social dissemination, $f + L + S$ all features. 86

5.4 Distribution of D^L values across growth and decline words, grouped by part of speech tag. * indicates $p < 0.05$ in one-tailed t-test between growth and decline D^L values. 87

5.5 Distribution of concordance scores (10-fold cross-validation) of the Cox regression models across feature sets. 89

7.1 Frequency and rate of verb integration over time, from 1% Twitter data sample from 2014-2019. 109

7.2 Top 10 artists in *SLA* and *USUK* categories. 117

7.3 Audience age distributions for the "youngest" and "oldest" *SLA* and *USUK* artists, queried from Facebook marketing API. 119

7.4 Unbalanced and balanced audience age distributions of *USUK* and *SLA* artists. The unbalanced age distribution exhibited a significant difference between the genres, while the balanced age distribution did not exhibit that difference. 120

7.5 Integrated verb use across social media text (blue) and newspaper text (orange). Each point represents a single word. 125

CHAPTER 1
INTRODUCTION

Language is often modeled as a self-contained system of generative constraints that exists to convey a speaker's own thoughts (Chomsky, 1986). However, communication does not simply entail transferring information efficiently between people, it also reflects a speaker's ability to navigate social life (Eggins, 2004). A person may claim to be *going to the park* among professional colleagues but later say *goin' to the park* with friends to signal informality and closeness (Labov, 2001). Sociolinguists often study language variation, or the alternation between competing variants (e.g. *going* vs. *goin'*), to reveal differences between social groups such as geographic dialects (Trudgill, 1974) and to investigate a speaker's communication goals in a given conversation (Auer, 2013). A person may moderate their language if they are speaking in front of an unfamiliar audience (Bell, 1984), signalling their attitude toward the topic under discussion (Preston, 2002), or showing their membership in a particular community (Eckert and McConnell-Ginet, 1992). Sociolinguists also study how long-term variation in language patterns over time leads to language *change*, as a particular variant spreads between communities or between people of different generations (Weinreich, Labov, and Herzog, 1968). These patterns are the basis for sociolinguistic theory that seeks to explain how people moderate their language use to create "social meaning" (Eckert, 2008).

In contrast to traditional sociolinguistics that has investigated spoken language variation (Labov, 2001), computational sociolinguistics (Nguyen et al., 2016) has worked to model large-scale language variation in *online* settings, particularly to address differences between populations such as geographic dialects (Eisenstein et al., 2010). For a better understanding of language variation, the internet hosts a variety of social media platforms where people leverage language choices to build community and express

themselves (Herring, 2012). Computational sociolinguistics research often focuses on differences in word usage, such as stylistic markers (e.g. hedges like *kinda*; Danescu-Niculescu-Mizil et al., 2012; Pavalanathan et al., 2017), and how these differences reflect known social systems such as gender (Bamman, Eisenstein, and Schnoebelen, 2014). Studies in computational sociolinguistics also rely on natural language processing techniques, including automatic language identification (Lui and Baldwin, 2012), to extract patterns of language variation that could otherwise be difficult to find manually, such as the use of different languages according to social context (Nguyen, Trieschnigg, and Cornips, 2015). This thesis builds on computational sociolinguistics by investigating open sociolinguistic questions that can benefit from the data available on social media.

Online communication platforms provide a domain to address questions that were often addressed with other approaches such as experiments (Lazer et al., 2009) due to their rare or complex nature. For example, sociolinguistic research often focuses on rare patterns of variation, such as adoption of new words which are relatively infrequent in spoken conversation. Instead of relying solely on self-reported data from participants, public discussions in online platforms can reveal natural social patterns such as a person's conversations with friends versus strangers (Fussell and Krauss, 1989). In contrast to traditional sociolinguistics studies that often are restricted to a fixed social context (Bucholtz, 1999), social media shows how speakers behave in multiple social contexts and therefore how language variation can serve speakers in different scenarios. This thesis seeks to leverage the social constructs in social media, such as audience and community structure, with the goal of testing open sociolinguistic questions that are otherwise difficult to address.

On top of these benefits for sociolinguistic research, investigating language variation online provides a broader understanding of how people leverage affordances in social platforms to achieve goals (Ren et al., 2011). Providing fine-grained insight into how

people navigate audience decisions in online discussions can inform writing affordances for platform users, similar to suggestions in how to help people navigate privacy decisions when sharing content (Acquisti, Brandimarte, and Loewenstein, 2015). For example, do people perceive an online "audience" as readily as an offline audience, and how readily is this reflected in their language choices (Marwick and boyd, 2011)? Understanding sociolinguistic variation in online discussions can also provide insight for community moderators who want to understand their members' participation in their community (Ling et al., 2005). An online community that is struggling to retain its members may better understand the needs of newcomers by testing whether the newcomers tend to adopt linguistic norms (Danescu-Niculescu-Mizil et al., 2013), and adjust the community's expectations of newcomers accordingly. As social media platforms become increasingly important for consuming information and connecting disparate populations (Schmidt et al., 2017), it will be equally important to determine how to apply sociolinguistic theory to better understand the overall experience of people on platforms.

Lastly, addressing sociolinguistic variation in the online world can help social computing researchers address more diverse populations (Wojcik and Hughes, 2019) who may be more difficult to study at scale, such as multilingual people. Traditional sociolinguistic research has focused on English and other widely-spoken languages at the expense of the "long tail" of less-studied varieties (Stanford, 2016), which are increasingly well-represented on the internet (Kim et al., 2014). Testing sociolinguistic questions in the context of a wide range of linguistic backgrounds can provide insight into how to adapt ideas such as "audience" to more complicated online situations, which in turn provides a critical lens to typical "Big Data" research (boyd and Crawford, 2012).

This thesis investigates language variation and change, with the goal of addressing open questions in sociolinguistics research. I leverage a variety of social media platforms, which provide a means to quantify complicated social constructs such as audience, to investigate both community-level language change and speaker-level differences in

language choice.

1.1 Thesis organization

The thesis is organized around three core research questions that traditional sociolinguistics work often struggles to answer, due to limited data on speakers' prior behavior and contexts of discussion, as well as limited data on global trends. I address these questions by applying natural language processing techniques to extract patterns of language variation from large-scale online discussion data and performing statistical analysis to assess the social factors that relate to language variation. The findings of these studies both provide answers to sociolinguistic inquiry and help inform the utility of social affordances offered by social media platforms, such as audience configurations.

1.1.1 Research question motivations

To maximize the utility of social media and NLP as a lens into language variation, I pose the following three research questions.

1. One central question for sociolinguistics is how speakers perceive other people in their conversation: how do speakers adjust to the assumed expectations of their listeners (Bell, 1984)? Participants may often accommodate to one another during the course of a conversation as a result of perceived social connection (Pardo, 2006). Furthermore, people who cannot view their listeners, such as radio broadcasters, often prepare for their imagined audience with an accent that matches the audience's assumed expectations (Coupland, 2001). While traditional sociolinguistics has studied this phenomenon in spoken communication, it not always easy to characterize the communication expectations in a given conversation and therefore difficult to predict a speaker's response in more complicated scenarios, such as when the speaker and audience are reacting to a shared experience unfolding in real time.

Fortunately, social media provides an ideal test bed for this question, as people in online discussions write messages to a variety of different types of audiences, some fully known and others partially known (Frobenius, 2014; Nguyen, Trieschnigg, and Cornips, 2015; Zhang et al., 2020). Furthermore, social media provides a more complete picture of conversational context and therefore reveals potential conversational expectations, such as the relative time of an ongoing event during which a speaker sends a message. The first open sociolinguistics question that I pose is therefore related to how speakers adjust to the expectations of partially-known conversation participants.

RQ1: How do speakers adjust their language to the assumed expectations of their community and their discussions, when they may not know the other participants?

2. Another key sociolinguistics question is the how language changes over time (Weinreich, Labov, and Herzog, 1968), especially how new words are adopted. Sociolinguists have focused heavily on disentangling the variety of *social* structures that can lead to change including demographics (e.g. young women seen as leaders; Labov, 2001), networks (weak ties help spread; Milroy and Milroy, 1985), and social identity (adopting change helps people construct their identity; Eckert, 2008). However, this kind of research is often limited by the scale of change that it can handle: since most forms of change such as new words are rare, sociolinguists often have trouble testing whether an observed pattern of change can *generalize* across different cases. While traditional studies help characterize particular types of change such as the Northern Cities vowel shift (McCarthy, 2011), they often struggle to "zoom out" and test factors that apply consistently across changes. Furthermore, since language change is not instantaneous but requires generations to propagate (Tagliamonte and D'Arcy, 2007), researchers often need to infer a change's progress from its adoption among people from different generations, e.g. if younger people have adopted a new form at a higher rate than old people.

Moving to social media to study word adoption provides a solution to both the issue of scale and time. First, monitoring public discussions on social media provides a window to a massive scale of data, which helps isolate rare cases of new words such as *bae* that spread through the online world (Grieve, Nini, and Guo, 2016; Kershaw, Rowe, and Stacey, 2016). The benefit of scale is particularly important when comparing multiple factors in language change, such as the relative importance of language-internal factors (structure) and external factors (social acceptance of change) (Metcalf, 2004): a new word may "succeed" both because it occurs in many linguistic contexts or because it is well-accepted in many social contexts. Such research require numerous simultaneous changes to test which factors *consistently* explain change. Second, public social media provides a long-term view that helps researchers identify cases of real, consistent language change as compared to ephemeral changes in word use (Kulkarni et al., 2015). I pose the following open question related to language change that social media can help address.

RQ2: How readily do linguistic context dissemination and social context dissemination explain the adoption of words in online communities?

3. The prior two questions considered primarily monolingual situations, where speakers navigate their listeners' expectations and participate in large-scale changes. Sociolinguists have also investigated *multilingual* speech decisions, such as code-switching between languages in the same sentence (e.g. *I start a sentence y termino en español*) (Poplack, 1980). In particular, a speaker's decision to choose one language over another relies partly on their *identity*, whether they use their language to signal affiliation to one culture over another (Auer, 2013; Gumperz, 1977). Often a speaker's identity manifests through the attitudes that they hold toward a particular social group, as when a multilingual speaker borrows words differently from another language if they hold a more or less positive view of the language's associated culture (Lev-Ari and Peperkamp, 2014). When assessing

multilingual speakers' motivations in language choice, traditional sociolinguistic studies often have difficulty identifying speaker attitudes in a naturalistic setting free from observer's bias (Cukor-Avila, 2000). Furthermore, while sociolinguists often focus deeply on the use of a multilingual speaker's language choices in a particular conversation or speech community (Androutsopoulos, 2007), typical spoken studies are not always able to test the same subject across different social contexts, which is necessary to test how consistently a speaker's attitudes affect both personal discussions and more public discussions.

To address these limitations, social media first provides a setting where people are expected to express their attitudes through participation in larger social movements (Gleason, 2013) and through sharing media that relates to particular cultures (Johnson and Ranzini, 2018). Second, social media can reveal how multilingual speakers make language choices in different language contexts, e.g. in front of different audiences (Nguyen, Trieschnigg, and Cornips, 2015), which is difficult to capture in spoken interview studies where a speaker is primarily talking to a limited group of people. As the third component of the thesis, I address the following research question in order to show the benefit of social media data in addressing sociolinguistic inquiry.

RQ3: For multilingual speakers, how consistently do social attitudes explain their choice of which language to use in online discussions?

1.1.2 Thesis studies

I address these research questions with the following set of studies.

1. RQ1: How do speakers adjust their language to the assumed expectations of their community and their discussions, when they may not know the other participants?

   (a) **Chapter 3**: First, I study the adoption of orthographic variant hashtags in a

community on Instagram where hashtag spelling is often manipulated to avoid content bans (e.g. *#anorexiaaa* from *#anorexia*). I find that the community-wide trend toward more extreme variants over time is driven by community newcomers, who later abandon these variant hashtags presumably to conform to the "older" members. This adds a new perspective on the typical "lifecycle" of community norm development (Danescu-Niculescu-Mizil et al., 2013), where change in the community is driven by newcomers who keep their initial language rather than abandoning it. Newcomers' behavior may therefore be driven by a perceived need to adapt to the community's presumably more "extreme" linguistic norms, which adds a constraint to the typical notion of how communities of practice develop norms (Lave and Wenger, 1991).

(b) **Chapter 4**: In this chapter, I propose a method to quantify collective attention in online discussions with the use of descriptor phrases (e.g. *San Juan, a city in Puerto Rico*). Within social media discussions related to major crisis events, authors tend to add descriptors in response to increased perceived audience needs (e.g. non-local audience) and remove descriptors for decreased audience needs (e.g. after the peak in post volume). By showing that people actively adapt to their audience in both static and dynamic contexts, I provide a new insight into the development of information status in public discussions (Prince, 1992) that shows how even strangers can converge on shared perceptions of events.

2. RQ2: How readily do linguistic context dissemination and social context dissemination explain the adoption of words in online communities?

   (a) **Chapter 5**: To address this question, I assess the relative importance of social and language factors in explaining long-term word adoption on Reddit, using a novel metric of linguistic dissemination to measure a word's tendency to

appear in multiple lexical contexts. Linguistic dissemination provides a more accurate prediction of when words will succeed *and* when they will fail, as compared to social factors. This finding demonstrates an important limitation to explaining language change through social evaluation alone (Altmann, Pierrehumbert, and Motter, 2011). This in turn helps address the larger question of language change by suggesting that a speaker's linguistic evaluation of a new word (how useful a word is) can outweigh the typical social evaluation (whether the word is socially acceptable).

3. RQ3: For multilingual speakers, how consistently do social attitudes explain their choice of which language to use in online discussions?

   (a) **Chapter 6**: To provide a strong basis for social attitude, I first study the choice between minority and majority languages in political discussion on Twitter, in the context of an independence referendum in Spain in 2017. I find that pro-independence people tend to use more of the minority language Catalan, that people tend to use more minority language when discussing the referendum, and that people use more majority language with a smaller audience, possibly to maximize the likelihood of engagement. This suggests that Catalan speakers tend to use the minority language for political stance in all situations, except in small-audience scenarios where eliciting a response (e.g. from a politician) is more important. The study advances typical notions of code-switching by demonstrating the political attitudes that underlie language choice, particularly as it can help minority-position people support their cause when they are addressing a broad, partially-known audience.

   (b) **Chapter 7**: Building on the political study, I turn to the adoption of loanwords into a target language as a potential scenario for social attitude expression. I investigate the use of *integrated* English loanwords in Spanish, a process by

which non-Spanish words gain native morphology (e.g. English *tweet* to Spanish verb *tuitear*). I quantify a speaker's cultural attitude with their degree of Latin American and US American music sharing on social media, and I investigate the relevance of *speaker-level* attributes on loanword integration. When predicting individual speakers' use of integrated versus other loanwords, I find that speakers' cultural attitudes explain integrated loanword use less well than demographic factors, such as language and geography. Taken together, these findings demonstrate that loanword integration is related more to macro-level language systems (i.e. formality) than to individual-level speaker choices, which reinforces prior findings about morphological integration being "instant" (Poplack and Dion, 2012). With respect to social attitude, the lack of a consistent media effect suggests that loanword integration does not have a strong tie to attitude expression as compared to the choice of language in the political referendum study.

Considering all the studies together, this thesis demonstrates that language variation in online discussions can help fill research gaps in sociolinguistics, by leveraging the large scale of social media, the more naturalistic setting for conversation observation, and the variety of social contexts in which speakers can participate. Rather than promoting a unified theory of communication, this thesis proposes situated approaches to understand the relevance of language variation to particular online spaces. Furthermore, the thesis provides a linguistically-informed lens for social computing researchers to better understand the possible motivations for platform members' behavior.

I provide a summary of the research questions and their connection to individual chapters in Table 1.1.

Table 1.1: Summary of the high-level research questions addressed by the thesis and the corresponding results.

Research question	Study results
RQ1: How do speakers adjust their language to the assumed expectations of their community and their discussions, when they may not know the other participants?	**Chapter 3**: Community newcomers introduce and later abandon the more "advanced" forms of hashtag variants, which drives the community trend toward more variation over time. **Chapter 4**: People responding to a crisis event adjust to their audience's assumed expectations, based on their own background and the event's dynamics.
RQ2: How readily do linguistic context dissemination and social context dissemination explain the adoption of words in online communities?	**Chapter 5**: Dissemination across linguistic contexts predicts word growth and decline more readily than dissemination across social contexts.
RQ3: For multilingual speakers, how consistently do social attitudes explain their choice of which language to use in online discussions?	**Chapter 6**: Multilingual speakers use the minority language in response to pro-independence attitudes and political discourse in general. **Chapter 7**: Speakers with different cultural attitudes do not use different forms of loanwords.

1.2 Thesis statement

This thesis **addresses open questions in sociolinguistics through quantitative analysis of language variation in online discussions, to address limitations in the current understanding of speakers' adaptation to listener expectations, the adoption of new words, and the language choices among multilingual speakers**.

1.3 Contributions of thesis

This thesis makes the following contributions.

1. **Addressing questions in sociolinguistics**

   The thesis finds that the language variation can provide insight into the role of communities of practice (Lave and Wenger, 1991), audience design (Bell, 1984), adoption of new words (Metcalf, 2004), and social attitudes (Ladegaard, 2000). The studies in this thesis leverage the large-scale nature of social media data to test language variation at the speaker-level across multiple social contexts, such as language choice among multilingual speakers in small-audience and large-audience contexts (Chapter 6). Rather than a single construct for all situations, this thesis reveals how different definitions for a construct, e.g. political versus cultural attitudes, can be useful to understand different types of social meaning as expressed through language. This thesis also compares the *tension* between different communicative factors as they present different mechanisms for speakers' language variation, such as the need for positive social evaluation (Altmann, Pierrehumbert, and Motter, 2011) versus the need for linguistic utility (Ito and Tagliamonte, 2003) in the context of adopting new words (Chapter 5).

   Furthermore, this insight connects language variation to broad patterns that emerge from online communication. The small-scale language choices that individuals make often lead to more complex system-level patterns, such as "waves" of

collective attention in response to breaking news events (Chapter 4). The affordances provided by social media platforms, such as explicit definitions of audience, can shape a person's social motivations and therefore their individual choices, including their participation in a community defined by hashtag use (Chapter 3). When studying large-scale language variation online, computational sociolinguistics researchers should consider a bottom-up view that connects concrete platform affordances to broad patterns of variation.

The growing field of computational sociolinguistics has begun to explore a wider range of variation (Nguyen et al., 2016), and this thesis focuses partly on variation in *structure*, such as the social meaning of differences in word form (Chapter 3, Chapter 7). Work from this thesis has encouraged similar explorations of variation in language structure in online spaces and beyond (Ndubuisi-Obi, Ghosh, and Jurgens, 2019; Hofmann, Pierrehumbert, and Schütze, 2020; Ryskina et al., 2020). Outside of the typical domain of social media, this thesis provides examples of language choices that deserve study in non-internet environments, e.g. newspaper coverage of current events (Staliūnaitė et al., 2018). On top of highlighting language structure, the results from my research can encourage researchers in the space of computational sociolinguistics to consider possible social *motivations* to explain variation in language online, rather than focusing solely on descriptive models (e.g. geographic modeling of language change). At a more fundamental level, this thesis pushes research in computational sociolinguistics to move beyond the *what* of language variation and investigate the *why*: i.e. what do patterns of change and variation tell us about possible speaker *goals* in online communication?

2. **Insight for social computing**

   On top of sociolinguistics insight, this thesis provides insight toward how affordances in social media platforms can shape members' behavior. In particular, I

offer a framework for understanding language variation as a window to how people participate in online discussions (Herring, 2012; Jaffe et al., 2012), particularly with respect to how speakers anticipate their audience and adapt to their communities. In addition to typical content analysis approaches e.g. word frequency modeling, social computing research should consider more fine-grained analysis of individual language choices such as style cues (Pavalanathan, Han, and Eisenstein, 2018), to help provide a richer speaker-level view of behavior online. The insight gained from sociolinguistic analysis can in turn help media platforms better anticipate the communication needs of platform authors, such as interventions that help writers preempt their audience's reaction (Zhang et al., 2020) or that allow moderators to track newcomer adaptation to norms in an online community (Hamilton et al., 2017).

The research approach in this thesis can encourage study into the more *nonstandard* side of language on social media such as slang, which is often normalized (Eisenstein, 2013b) or ignored in favor of standard approaches such as lexicon matching. This thesis argues for a more situated approach to sociolinguistic analysis (Patton et al., 2020), focusing on the specifics of a given community or online space to address patterns of participation in discussions. Taking a situated approach to language variation in the context of social computing can also yield practical takeaways for fields that rely on social computing, such as crisis informatics (Chapter 4) and political science (Chapter 6). Social media systems do not exist in a vacuum, and stakeholders, including public opinion monitors, often use the systems to gain a better understanding of "on-the-ground" reactions to events. Language variation represents an important component of public opinion that can inform stakeholders' decisions. As social computing platforms become more specialized and pervasive, it will be important to consider a wide range of communication strategies that emerge on these platforms, including systematic

variation in language.

3. **Extensible linguistic analysis methods**

   This thesis proposes several unique approaches to natural language processing, such as a method for location descriptor detection (Chapter 4) and a metric for context dissemination (Chapter 5). The methods detailed in this thesis will help similar sociolinguistic research scale up its approach and address more fine-grained linguistic phenomena that are difficult to identify manually. Researchers will also benefit from the fact that the methods are interpretable and readily implemented without significant overhead in terms of machine learning. Implementation is especially important in scenarios where ground-truth labels for content analysis are unavailable or where data is generally sparse. Furthermore, the methods in this thesis provide a benchmark that can be extended with other NLP methods for further improvement, e.g. augmenting the notion of linguistic dissemination (Chapter 5) with semantic word representations.

CHAPTER 2
BACKGROUND

In this chapter I review the necessary background for the thesis.

I begin with an overview of language variation (§ 2.1) and the thesis's focus on variation in structure (§ 2.1.1). Next, I provide a summary of the relevant sociolinguistic theory that supports the studies in speaker-level language variation (§ 2.1.2) and language change (§ 2.1.3). Lastly, I connect the work of this thesis with a broader line of work investigating language variation on the internet (§ 2.2), focusing on computational sociolinguistics (§ 2.2.1) and social computing (§ 2.2.2).

2.1 Language variation and change

The language that people use in everyday conversation reflects societal norms and expectations, which includes the identity that people present to others (Goffman, 1978), the relationships that people navigate (Granovetter, 1973), and the communities to which people belong (Lave and Wenger, 1991). A person's dominant dialect, which includes unique pronunciation, words and grammar, often reflects the region in which they were raised or with which they most strongly identify (Chambers and Trudgill, 1998): for example, people in the United States variously use the words *soda*, *pop* and *coke* to mean a carbonated beverage, depending on where they grew up (Katz, 2016). In the course of a conversation, people often adopt the style of their conversation partners as a signal of accommodation to their viewpoint (Giles, Coupland, and Coupland, 1991). After joining a speech community, people may bring in new styles of speaking that spread to other members of the community, thereby becoming a new communication norm (Trudgill, 1974). Additionally, a language difference may be *transmitted* between generations, as younger speakers often acquire the language patterns of older speakers imperfectly,

resulting in a long-term change over several generations (Labov, 2007).

This kind of work is the study of *variationist* sociolinguistics, which studies how language varies between speakers and changes over time (Labov, 1972). Variationist sociolinguists focuses on explaining alternation in a particular linguistic *variable*, such as the alternation between *-ing* and *-in'* in verbs (*going* / *goin'*), by systematic investigation of social structures and motivations. Traditional sociolinguistic work investigated pronunciation differences among social groups, including regional vowel shifts like the Northern Cities Shift (McCarthy, 2011), partly because accent differences are highly perceptible among most native speakers (Long and Preston, 2002). The typical methods involved include participant interviews (D'Arcy and Tagliamonte, 2015), long-term participant observation (Eckert, 1989), and experiments such as the matched guise test, where speakers make identity judgments about a hidden speaker based only on their language (Preston, 2002). Research in the field initially focused on socioeconomic class to explain language variation (Labov, 1963; Labov, 2006) but has since expanded to a wide range of factors including race (Wolfram and Thomas, 2008), community structure (Milroy and Milroy, 1985) and persona (Bucholtz, 1999). In addition to providing theoretical insight into social processes underlying language variation and change, sociolinguistic research has also provided descriptive insight into the many dialects of English, such as the geographic dialects of American English (Labov, Ash, and Boberg, 2008). This descriptive work has helped to legitimize minority language varieties in the face of social stigma, notably in the case of African American English as a contested dialect in the United States (Perry and Delpit, 1998).

2.1.1 Variation in structure

This thesis primarily studies variation in linguistic structure, which includes word structure (Brody and Diakopoulos, 2011) and sentence structure (Staliūnaitė et al., 2018). I provide examples of each type of variation here.

- **Word structure**: the addition of extra characters or morphemes to a word to mark it as different from the "standard" form: e.g. *lollll* as an extension from *lol*.
- **Sentence structure**: the addition of syntactic phrases (e.g. dependent clause) to a given word: e.g. *Atlanta, a city in the state of Georgia*.

Variationist sociolinguistic research has studied language structure as it relates to social differentiation, notably structure of loanwords (Poplack, Sankoff, and Miller, 1988) and syntax (Blake, 1997). For syntax, sociolinguists have focused on variables such as null copula (e.g. *he eating* versus *he is eating*) which are trademarks of dialects such as African American English (Wolfram and Thomas, 2008), to present a more full descriptive picture of dialects outside of pronunciation and word choice (Zanuttini et al., 2018). Variation in language structure may appear non-systematic to speakers of standard language varieties, e.g. the double negative of *I didn't do nothing*, but sociolinguistic research has repeatedly demonstrated that such variation has consistent linguistic constraints and social meaning (Nevalainen, 2006). Similarly, a multilingual speaker may change how they adopt loanwords in conversation, e.g. using a phonetic structure that is closer to their native language (pronouncing English *tweet* with "correct" Spanish phonology *tuit*), in accordance with their prior knowledge of the source language (Auer, 2013).

While frequently studied in the spoken domain, this kind of variation has only recently begun to be investigated in the written domain of online communication, despite the possibility for structure to have a large impact on how dialects are represented on social media through syntactic variables (Blodgett, Wei, and O'Connor, 2018). The lack of research stems partly from the methods: research into variation in structure requires a robust scheme for identifying the different forms of structure, which can require specialized text processing methods. Manually counting variation such as null copulas is more complicated than identifying variation in pronunciation and word use (Rickford et al., 1991), and therefore difficult to scale beyond a few speakers. This thesis leverages

NLP methods such as parsing and edit distance to quantify variation in structure with high precision at scale. Furthermore, researchers have rarely considered how different cases of structural sociolinguistic variation can reflect similar social constraints, such as how different types of syntactic variation can reflect common patterns of dialect development (Kortmann and Szmrecsanyi, 2011). This thesis provides evidence that different patterns of variation such as phrase syntax and word structure can be explained by common underlying social constraints in online written discussions.

I now turn to the relevant theory from sociolinguistics that drives the research in this thesis, namely speaker-level variation and change at the level of language.

2.1.2 Sociolinguistic theory: speaker-level motivations

Language variation can shed light on how people create social meaning through their behavior in conversation (Eckert and Rickford, 2001). Even a decision as simple as expressing laughter as *lol* instead of *haha* can reflect the intention to be informal with one's peers (Tagliamonte and Denis, 2008). This thesis pursues open questions in sociolinguistics research by turning to online discussions, where familiar constructs such as "community" may be different than their offline counterparts.

In this section I review the social theory relevant to the studies in this thesis: conversation adaptation (Chapter 4), communities of practice (Chapter 3), and social attitudes (Chapter 6, Chapter 7).

Conversation adaptation

People bring different expectations of others to a conversation, and their expectations to the other speakers generally adhere to the assumption that the conversation follows rational rules (Grice, 1975). A speaker may change their language use and *accommodate* to others in order to shape their evolving relationship with their listeners (Ladegaard, 1995) or to highlight different parts of their message (Galati and Brennan, 2010).

Speakers with different backgrounds have been shown to converge to similar pronunciations during conversation (Pardo et al., 2012), which may be conditioned on their prior attitude toward the other speakers (Babel, 2010). Furthermore, the development of a shared "common ground" among conversation participants, such as a shared experience (Doyle and Frank, 2015), can cause convergence to identical speech norms due to speakers' similar intentions. In a study of a politician's speeches, Acton and Potts argue that the speaker's preference for *that* as a demonstrative (e.g. *that new commitment* vs. *the new commitment*) reflects the speaker's intention to reference a common ground implicitly developed with the listeners (Acton and Potts, 2014).

One especially notable aspect of conversation adaptation is the idea of *audience* design, where a person anticipates their audience before speaking and adjusts their language accordingly (Bell, 1984). For instance, Coupland showed that radio announcers in Wales intentionally adjust to their audience through active use of Welsh English dialect features (Coupland, 2001). Audience design applies to situations where a speaker has to address a partially-known or unknown set of listeners, such as radio broadcasts and video blogs (Bell, 1991a; Frobenius, 2014). In the face of a highly uncertain audience, a speaker may err on the side of greater politeness and use more formal language variants (Brown, Levinson, and Levinson, 1987). In the context of social media, Pavalanathan and Eisenstein tested the language style of Twitter authors and showed that they adopt more formal language when using hashtags, which implicitly reach a larger audience (Pavalanathan and Eisenstein, 2015a). Similarly, multilingual authors often choose between languages to best suit their intended audience in reply-threads (Nguyen, Trieschnigg, and Cornips, 2015; Shoemark, Kirby, and Goldwater, 2017), which may mean more minority language use for close friends.

This thesis leverages social media to investigate how language variation can be explained by consistent conversational expectations, including how people adapt to a partially-known audience that develops during a breaking news event (Chapter 4) and how

people anticipate the expectations of other discussion participants in political discourse (Chapter 6).

Communities of practice

In addition to adapting to a particular conversation, speakers are often expected to adapt to the expectations of a broader *speech community* when choosing between language variants (Holmes and Meyerhoff, 1999). From a coarse-grained perspective, a speech community may encompass the complete population of speakers of a particular language, e.g. all speakers of French are considered part of the "Francophone" community (Salhi, 2002). Variationist sociolinguistics has traditionally defined speech communities through geography, e.g. the region in which a person grew up and acquired their native language variety (Gumperz, 2009), or through demographics, e.g. communities based on speaker race (Green, 2002). As sociolinguistics has expanded its study of the social construction of identity, researchers have expanded their notion of speech community to include more flexible groups in which different aspects of identity can intersect (Eckert, 2012), such as sub-cultures that emerge among high schoolers (Bucholtz, 1999). In particular, sociolinguists of the "third wave" have adopted the model of *communities of practice* (Eckert and McConnell-Ginet, 1992) to understand how language variation helps speakers construct their identity in tandem with community goals.

A community of practice is a group of people who share a set of goals, and who demonstrate their expertise with respect to such goals through the development of consistent practices (Dubé, Bourhis, and Jacob, 2006). Through the process of Legitimate Peripheral Participation (LPP), a community newcomer adopts the shared practices by moving from the periphery to the center of the community (Lave and Wenger, 1991). New Wikipedia editors often learn how to edit articles from explicit advice and passive observation of more experienced editors, thereby joining an online community of practice (Bryant, Forte, and Bruckman, 2005). Community norms often receive

enforcement from members with more authority or experience (Blashki and Nichol, 2005), such as regular and well-connected members (Kooti et al., 2012b). However, other studies have shown that newcomers are early adopters of ongoing changes; these individuals then become conservative, maintaining the practices that were innovative at the time when they joined the community (Danescu-Niculescu-Mizil et al., 2013). Community practices may also be adopted differently depending on member lifespan: for instance, transient, or short-lifespan, members often invest less in community practices than committed, or long-lifespan, members (Ren et al., 2011). The development of norms, particularly linguistic norms, in an online community reflects a commitment to shared goals and values, e.g. in contrast to ephemeral communities that do not develop a sufficiently committed base to establish clear norms (Cunha et al., 2019). In the case of minority communities, a commitment to "hidden" language norms such as misspelled hashtags reflects the intention to avoid detection (Chancellor et al., 2016).

This thesis leverages LPP in the context of an online discussion community and finds that the community-wide adoption of linguistic variants tends to be driven by committed newcomers (Chapter 3).

Social attitudes

Moving from the perspective of other conversation participants to the speaker themselves, a speaker may choose between language variants based on *social attitudes*. Here I consider a social attitude as any positive or negative evaluation that a person expresses regarding another social group or institution to which the person does or does not belong (Olson and Zanna, 1993; Preston, 2002). While abstract, social attitudes are often manifested in language variation as people implicitly or explicitly evaluate the social group associated with a particular variety (Ladegaard, 2000). A US politician who holds typically conservative beliefs may implicitly signal their attitude toward a foreign entity by pronouncing the entity's name with a more American accent (Hall-Lew, Coppock, and

Starr, 2010). Social attitudes may reflect stereotypes rather than fact or personal experience with another group (Ferrara and Bell, 1995), often related to an exaggerated perception of language behavior such as the "Valley Girl" accent of California American English (Dailey-O'Cain, 2000). Furthermore, social attitudes form an important part of a speaker's broader sociolinguistic *identity* (Bucholtz and Hall, 2005), which is constructed dynamically during conversation using a variety of linguistic resources. In a formative study of language variation among young people, Bucholtz showed that young women in high school who identified as nerds avoided language patterns that were standard among other high schoolers (e.g. new slang and perceived-incorrect syntax) to express their negative attitudes toward expectations of normal behavior (Bucholtz, 1999). A person's social attitudes are more context-dependent and flexible than other aspects of identity such as demographics or social status (Snell, 2010).

Social attitudes represent an important part of multilingual speaker behavior in cases such as code-switching (Jaffe, 2007) and loanword adoption (Zenner, Speelman, and Geeraerts, 2015), where language choice may relate to the speaker's beliefs about the relative status of the different languages. The role of attitudes is especially clear in situations where the languages have explicit social value, such as when a majority language is treated as more prestigious (e.g. English versus Spanish) (Lipski, 2005). In political discussions, a speaker's use of a language that is tied to a political cause such as independence can signal their affiliation with the cause (Shoemark et al., 2017), as in the case of Belgium which has a strong connection between regional politics and language use (Blommaert, 2011). Furthermore, language choice may be tied to broader attitudes about specific cultures, as when multilingual immigrants use language to index attitudes toward the culture from which they are separated (Christiansen, 2015; Low, Sarkar, and Winer, 2009). Investigating multilingual behavior on an immigrant web forum, Androutsopoulos finds that code-switching to native language helps participants express their commitment to aspects of their heritage culture (Androutsopoulos, 2007).

The longitudinal nature of social media data provides an opportunity to investigate the role of naturally-expressed attitudes at scale and with less risk of observer's bias. To leverage this affordance, this thesis explores the relationship of a person's presumed political and cultural attitudes with their language choice (Chapter 6) and loanword use (Chapter 7). This thesis focuses on the *behavioral* aspect of attitude (Triandis, 1991), i.e. how people act toward particular targets, rather than affect or cognition (Olson and Zanna, 1993) since social media does not reliably reveal all aspects of a person's mental state.

2.1.3 Sociolinguistic theory: language change

In addition to language variation at the speaker level, this thesis also seeks to leverage social media to address questions in systemic language change that are otherwise hard to address through spoken data. Here, I review relevant prior sociolinguistic work that informs the research question related to long-term language change (Chapter 5).

Word adoption

One of the most basic forms of language change is word adoption, by which a word enters a speech community's lexicon (Chesley and Baayen, 2010; Pierrehumbert, 2012). In monolingual settings, new words may arise through the mutation of existing forms by processes such as truncation (e.g, *favorite* to *fave*; Grieve, Nini, and Guo, 2016) and blending (e.g., *web+log* to *weblog* to *blog*; Cook, 2010). In addition, the internet has fostered an abundance of new words to reduce writing time and replicate spoken conventions, such as acronyms (*lol*) and onomatopoeia (*ugh*) (McCulloch, 2019). In multilingual settings, a new word may be borrowed as a loanword from another language (Chesley and Baayen, 2010), often from a majority language like English, to fill cultural or technological gaps (Zenner, Speelman, and Geeraerts, 2012) (e.g. Spanish verb *googlear* from the English verb *Google*). Sociolinguists have been careful to distinguish "nonce" words (Poplack and Dion, 2012) that appear within a single utterance or

conversation from fully adopted words, since the former represents a conversation-level change rather than a change in the language.

For spoken language, sociolinguists have traditionally studied word adoption by comparing adoption rates across age groups, under the "apparent time" assumption that younger speakers' language reflects changes in progress (Labov, 1963; Tagliamonte and D'Arcy, 2007). The intersection of age and other demographic attributes has proven useful as well, as younger women were traditionally considered to be the most consistent drivers of language change (Labov, 1990). In addition, sociolinguists have also considered network approaches whereby words diffuse between speech communities through weak social ties between speakers (Kerswill and Williams, 2000; Milroy and Milroy, 1985). Studying social networks in language change has helped compare the role of central members versus peripheral members of a speech community in sound changes (Fagyal et al., 2010), as the latter are often responsible for incoming change. In contrast these more structured approaches to language change, modern sociolinguists often frame word adoption as a question of *indexicality*, meaning that speakers adopt new linguistic norms to index social identity e.g. geek vs. jock (Eckert, 2012; Rickford and Price, 2013). In a study of Pittsburgh English speakers, Johnstone, Andrus, and Danielson found that the adoption of local words such as *yinz* (plural *you*) coincided with speakers' growing awareness and construction of local "Pittsburghese" identity (Johnstone, Andrus, and Danielson, 2006). This indexical approach proposes that language change stems from stylistic choices in conversation related to self-presentation (Bucholtz, 1999; Goffman, 1978), and therefore that language change is an inevitable product of socialization (Eckert, 2016).

To study word adoption in written language, sociolinguists have leveraged large-scale diachronic corpora of newspapers and books to track word adoption, which is often used as a proxy for social or technological change (Davies, 2014; Juola, 2013). The proliferation of discussions on social media and online forums has provided an ample

source of data for word adoption studies (Goel et al., 2016; Grieve, Nini, and Guo, 2016) and has illuminated rare types of word adoption that would otherwise be difficult to find in e.g. books or interviews. For instance, novel lexical blends such as *stan* (*stalker* + *fan*) are somewhat rare in spoken conversation but easy to study with more dense social media data (Cook, 2012).

Dissemination in word adoption

A variety of factors have been proposed to explain why some words take off and others are ignored (Metcalf, 2004), including surface-level features such as spelling (Kershaw, Rowe, and Stacey, 2016) and context-level *dissemination* (Altmann, Pierrehumbert, and Motter, 2011). A word that is widely adopted among a variety of people or groups, i.e. that is *socially* disseminated, may be seen as positively evaluated by social groups and therefore useful to adopt as a marker of social capital (Garley and Hockenmaier, 2012; Tredici and Fernández, 2018). In addition, a word that can be used in a variety of different lexical contexts, i.e. that is *linguistically* disseminated, may prove useful to speakers and out-compete more "rigid" word choices (Ito and Tagliamonte, 2003; Partington, 1993). This is a form of language structure (syntax) that provides insight into the relative utility of a word, which plays a part of the larger word adoption lifecycle. While important for different reasons, these types of word-success factors are not often systematically compared in large-scale settings.

In Chapter 5, I compare the relative importance of these dissemination factors in long-term growth and decline of nonstandard words online.

2.2 Language variation on the internet

While addressing open questions in sociolinguistics in general, this thesis joins a larger conversation about what language variation can tell us about social behavior on the internet. In this section, I review how the thesis connects with recent work in

computational sociolinguistics and in social computing.

2.2.1 Computational sociolinguistics

The field of sociolinguistics has recently turned toward computational methods, with the goal of leveraging the highly structured social data available in online and other written communication to explain patterns of langauge variation (Nguyen et al., 2016). In contrast to other work in natural language processing, computational sociolinguistics seeks to explain language use through a speaker's social background rather than through linguistic context alone. Researchers have found that geographic language variation in social media reflects known dialect differences (Eisenstein, 2013a; Kulkarni, Perozzi, and Skiena, 2016), such as the use of slang word *hella* by California Twitter users. In addition, gender differences and age differences in language use are often reproduced on social media (Bamman, Eisenstein, and Schnoebelen, 2014; Nguyen et al., 2013), such as a tendency for female-associated people to use more social and emotional language. Models to explain language variation include generative models (Eisenstein et al., 2010; Blodgett, Green, and O'Connor, 2016) and word embedding models, which combine social and linguistic signals to improve accuracy in downstream tasks (Bamman, Dyer, and Smith, 2014; Garimella, Banea, and Mihalcea, 2017).

In addition to identifying differences between social groups, computational sociolinguistics has expanded the scope of how researchers can study processes of language change. The availability of more dense social network data provides a means to quantify tie strength and thereby predict the spread of new words in online communities (Goel et al., 2016). Furthermore, researchers can also investigate patterns of change within specific online communities (Danescu-Niculescu-Mizil et al., 2013), which may be much faster and more related to community membership status than typical long-term changes (Tan and Lee, 2015). To explain language change, researchers have proposed a variety of models and metrics including autoregressive models (Eisenstein

et al., 2014), divergence between language models trained at different time periods (Zhang et al., 2017), and semantic differences between word embeddings (Kulkarni et al., 2015).

Outside of testing hypotheses about language variation, work in computational sociolinguistics can contribute to downstream applications. For one, understanding language differences among social groups has led to discoveries about NLP systems' bias against certain language patterns, particularly African American English with respect to parsing (Blodgett, Wei, and O'Connor, 2018) and toxic comment detection (Sap et al., 2019). In the same vein, studying language differences among social groups can help researchers develop personalized NLP models that adapt to an author's language (Del Tredici et al., 2019; Yang and Eisenstein, 2017). In terms of change, modeling language as a dynamic system can help NLP models adapt automatically to previously unseen time periods (Yang and Eisenstein, 2015; Huang and Paul, 2018), which is especially important in cases of semantic change among words.

Typically work in computational sociolinguistics has focused on word-level variation, due in part to the NLP perspective of treating words as atomic units. While useful, this ignores a wide swath of non-word linguistic phenomena that can reveal insight into social processes, such as variation in syntax due to age (Johannsen, Hovy, and Søgaard, 2015). This thesis expands the scope of computational sociolinguistics to study variation in structure, focusing on linguistic phenomena that can be detected with high precision to guarantee construct validity.

2.2.2 Social computing

Social computing platforms such as Twitter and Facebook were developed as a means for sharing information and maintaining relationships online (boyd and Ellison, 2007). While many such sites were initially developed within the U.S., social media platforms have been adopted around the world, thanks in part to increased mobile phone access (Poushter, Bishop, and Chwe, 2018). Researchers have generally studied the public-facing

discussions on social media platforms as opposed to private conversations, due to ease of access and reduced ethical considerations (Metcalf and Crawford, 2016). Observational studies of social media provide a useful alternative to experimental studies, which can assess causal relationships but also have a larger chance of observer bias (Wang et al., 2019a). The diversity of discussions on social media platforms has informed the study of a variety of social sciences, including political science (Sauter and Bruns, 2015), public health (Paul and Dredze, 2011), and journalism (Flew et al., 2012).

Research in social computing systems often focuses on what language use can reveal about the design and social structure of online discussion spaces (Bucher and Helmond, 2017). If a platform institutes a large-scale change such as Twitter's shift to 280 characters, people often adapt their language to the new restrictions and affordances (Gligorić, Anderson, and West, 2018; Pavalanathan and Eisenstein, 2016). Understanding how people modulate their writing can reveal the communication goals that authors have within their platforms, such as maximizing the spread of their message using highly charged language (Brady et al., 2017). Content moderators rely on reports of undesirable behavior (e.g. hate speech) to maintain community standards, and such behavior can be quantified through language choices (Chandrasekharan et al., 2017; Pavalanathan, Han, and Eisenstein, 2018). Since most discussion platforms rely on textual communication instead of e.g. voice communication (Jiang et al., 2019), social computing research often focuses on word usage to quantify linguistic patterns and behavior more generally.

In this thesis, I focus on several areas where sociolinguistic analysis can contribute to social computing research, including information dissemination (Chapter 4, Chapter 5) and conversation dynamics (Chapter 3, Chapter 4). With respect to information dissemination, new word forms can correlate with emergent forms of communication that reveal the needs of platform users, such as the widespread adoption of the retweet convention on Twitter (Kooti et al., 2012b). Sociolinguistics provides a framework for

understanding the emergence of such conventions as a competition among likely variants (Ito and Tagliamonte, 2003; Tagliamonte and Smith, 2006) and for comparing the relative influence of different social and communicative constraints on such variants. Furthermore, identifying leaders of linguistic change can also reveal the importance of their relative social position in a particular online space (Blythe and Croft, 2012; Stuart-Smith and Timmins, 2010). People who tend to adopt a language change before others may be more generally "ahead of the curve" with respect to community norms in general, such as social behaviors that others have not yet picked up.

With respect to conversation dynamics, social computing platforms often spend significant time and effort to understand the health of their communities (Kim, 2006), often addressed through active moderation. One dynamic that often requires attention is the turnover in community membership, since a community that loses newcomers before they fully commit may not have a long-term future (Althoff and Leskovec, 2015; Yang et al., 2010). Newcomers who explicitly adopt the community's language patterns are seen as accommodating to the community and more likely to remain long-term (Danescu-Niculescu-Mizil et al., 2013; Hamilton et al., 2017; Tan and Lee, 2015), while communities that make it difficult for newcomers to learn their norms may see more turnover (August et al., 2020; Nguyen and Rose, 2011). Platforms must additionally consider how readily their platforms can enable members to spread information quickly and in useful ways (Kogan, Palen, and Anderson, 2015), when confronted with fast-moving events. While a platform like Twitter enables rapid sharing of useful information through affordances such as shares and hashtags, it also can enable coordinated disinformation campaigns (Michael, 2017; Stewart et al., 2017b), which means that platform designers must work to support rational communication needs without helping bad actors.

This thesis provides linguistic insight into how people conform to others' expectations (RQ1), help to spread innovations (RQ2), and express their own identity

(RQ3) in online discussions. The findings of these studies can help social computing researchers understand the affordances of online platforms that allow people to communicate even amid strangers.

CHAPTER 3

ADOPTION OF VARIANT HASHTAGS IN ONLINE COMMUNITIES

Speakers are expected to make adjustments for their listeners in a conversation, such as switching to more formal speech (*goin'* to *going*) when among people with whom they are less socially close. However, while sociolinguists have studied such accommodations in structured settings such as radio broadcasts, it is less well understood how speakers adjust to in more uncertain situations, such as when reacting to an ongoing event with unknown listeners. The first part of this thesis (RQ1) investigates how people adjust their language use to conversation expectations when their audience is only partly known.

For the first study in this part of the thesis, I consider community-level change in hashtag use in a particular group on Instagram. In unstructured platforms like Instagram, hashtags can be used to organize social movements (Gleason, 2013) and communities that share common interests (Sauter and Bruns, 2015). The spelling of hashtags may be modified to emphasize different parts of the message (Tsur and Rappoport, 2015) and to evade content blocking (Chancellor et al., 2016). While such changes have been studied in aggregate, it is less well-understood how these changes may relate to speaker-level goals such as the intent to join a community with which the hashtags are affiliated.

Using hashtag spelling as an example of language variation, I investigate a hidden community on Instagram that used modified hashtags to avoid a content ban. I show that the community-wide trend toward more variation over time is driven mainly by newcomers who are especially committed, and that "deeper" variation correlates with social engagement. This adds new insight to the typical community of practice model by which newcomers keep their initial practices and thereby slowly change the community norms as old members are phased out (Danescu-Niculescu-Mizil et al., 2013). In contrast, the newcomers actually abandon their initial practices even while pushing the

community's practices "forward."

Note: Content for this chapter is drawn from Stewart et al. (2017a). This work was completed with the help of Stevie Chancellor, Munmun De Choudhury, and Jacob Eisenstein.

3.1 Motivation

Online communities are defined by their membership and the shared practices of their members. The adoption of such practices can differentiate new members from regular community members, as new members must learn the community's practices in order to be considered a regular community participant (Bryant, Forte, and Bruckman, 2005; Lave and Wenger, 1991). Among community practices, language plays a particularly important role as a signal of shared identity (Labov, 2001). In the online setting, nonstandard **orthography** such as "leet speek" can differentiate community newcomers from accepted members (Androutsopoulos, 2011). As important as language practices are, they are subject to constant change as a result of exogenous and endogenous events (Danescu-Niculescu-Mizil et al., 2013; Kooti et al., 2012b).

Who in a community drives these changes? If changing practices are not adopted by all community members, then what characterizes the members who accept and advance these changes? The social meaning of language change in online communities can be better understood by linking language change to membership dynamics, i.e., the progression of individual community members from new to regular member. For example, studies have shown that the adoption of slang words and jargon follows predictable temporal patterns, both at the community level and over the lifespan of individual community members (Danescu-Niculescu-Mizil et al., 2013; Eisenstein et al., 2014). This lifecycle pattern mirrors the generational aspect of language change by which children acquire a dialect from their parents and peers, and then retain the dialect into adulthood (the adult language stability assumption; Labov, 2001).

However, language change may also result from exogenous shocks, such as a content ban in an online community (Chancellor et al., 2016; Hiruncharoenvate, Lin, and Gilbert, 2015). In 2012, Instagram banned hashtags that promoted eating disorder behaviors, or "pro-ED" content, such as *#thinspo* (Hasan, 2012). In response, members of the pro-ED community adopted orthographic variations of hashtags to circumvent the ban. Over time, these hashtags grew more popular and more complex, becoming increasingly distant from the original spellings.

This chapter addresses how speakers choose language variants in response to the expectations of a partially-known community (RQ1). To address this question in the context of the banned hashtags, I adopted a novel approach to measure change in community practices via orthographic variation, exploring the following three research questions:

- *RQ1*: Who uses orthographic variants?
- *RQ2*: Is depth of variation affected by membership attributes (i.e. age and lifespan)?
- *RQ3*: Does orthographic variation affect social reception (via likes and comments) of pro-ED content?

I first addressed the correlation of orthographic variation to the behavior of pro-ED community members and then the social reception of such variation. In RQs 1 and 2, I focused on two variables that define community membership: *age* in the community and *lifespan*. Prior work has highlighted the role of member age as a factor in the adoption of practices: newcomers can drive adoption of new words within a community but may become more resistant to change as they spend more time in the community (Danescu-Niculescu-Mizil et al., 2013). Furthermore, member lifespan, or total duration of time spent in the community, can impact adoption of community practices (Ren et al., 2011). RQ3 addressed the social relevance of orthographic variation, which can help explain its adoption within the community.

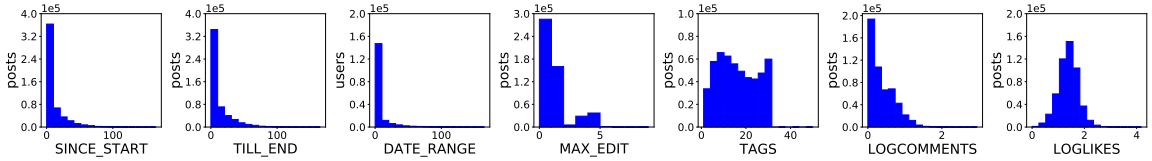


Figure 3.1: Summary histograms for all variables of interest, including relative time (e.g. DATE_RANGE), linguistic (MAX_EDIT) and social variables (LOGCOMMENTS).

To address these questions, I analyzed over two million Instagram posts and nearly 700 orthographic variants of pro-ED hashtags on Instagram. I found that in this community, orthographic variation is driven primarily by newcomers, especially those who will become long-term participants: these individuals were more likely to use orthographic variants, particularly deep variants that are far from the original spellings. The depth of orthographic variation was also correlated with community engagement: messages containing deeper orthographic variants received more likes.

3.2 Data

I employed a dataset with over two million Instagram posts, gathered from a set of "pro-ED" hashtags which promote disordered eating and exercise behaviors (Chancellor et al., 2016). This dataset includes manual annotations of the links between hundreds of orthographic variant hashtags and their original spellings.

Knowing the original spelling for each variant (e.g., that *#anarexyia* is related to *#anorexia*) makes it possible to compute the distance between the variant and source, and thus to quantify the depth of variation.

3.2.1 Data collection

Details of data collection can be found in the original paper by Chancellor et al. (2016). I summarize below only the most relevant points for this research.

The dataset was acquired in late 2014, using the public Instagram API to search for pro-ED hashtags. Because many hashtags could not be queried directly due to the

Table 3.1: Summary of orthographic variants grouped by edit distance. The edit distance 1 group has the greatest variety of source hashtags and unique variants, while the edit distance 4 group has the lowest variety. This study is restricted to variant hashtags with edit distance at or below 4, due to data sparsity above edit distance 4.

Edit distance	Top 3 variants	Source hashtags	Unique variants	% posts with at least one variant
1	*anarexia, bulimic, eatingdisorders*	17	253	41.1%
2	*anarexyia, thinspooo, thynspoo*	15	221	2.07%
3	*secretsociety123, thinspoooo, thygap*	15	108	9.60%
4	*secret_society123, secretsociety_123, thinspooooo*	10	50	10.4%

Instagram bans, Chancellor *et al.* identified a set of nine non-banned "seed tags" related to eating disorders. They gathered posts on those seed hashtags for 30 days, and identified the 222 most popular hashtags related to pro-ED behaviors. They manually removed hashtags that were ambiguous (e.g. *#fat*) or related to eating disorder recovery (e.g. *#anorexiarecovery*). This resulted in a set of 72 hashtags, which they used to gather a large dataset. After removing posts with recovery hashtags, the dataset contained 6.5 million posts, dating between January 2011 and November 2014.

From these 6.5 million posts, Chancellor *et al.* manually checked the top 200 most popular hashtags to see how many were banned by Instagram or placed on a "content advisory" (Hasan, 2012). They found seventeen source hashtags (e.g. *#thighgap* or *#anorexia*) that underwent some form of Instagram intervention. They then developed a set of regular expressions (e.g. $an * a*$ for *#ana*) to extract semantically similar yet orthographically variant hashtags from the source hashtags. The manual rating yielded 672 unique *orthographic variants*, and seventeen source hashtags, totaling 689 hashtags, which I study here.

In total, the dataset has 2,416,259 posts from January 2011 to November 2014, each of which contains at least one orthographic variant or source hashtag. Of these, 51% contain at least one variant and no source hashtags.

Qualitatively, the variant hashtags showed some systematic patterns in how they are modified. Vowels would often be replaced with similar sounding vowels (*anorexya* from

anorexia), and consonants with similar sounding consonants (*thighcap* from *thighgap*). Characters that are neighbors on QWERTY keyboards were frequently substituted for one another (*thinsporation* from *thinspiration*). Furthermore, some variants showed consistent addition of "throwaway" characters that are easily ignored to reveal the original hashtag (*secretsociety_123* from *secretsociety*). These systematic patterns suggested that the users who created these variants tried to minimize the amount of work that other users would need to "decode" the variants.

3.2.2 Feature extraction

The following features were extracted from each post and associated Instagram community member:

- Real time of post, measured in weeks since Instagram instituted a ban on several pro-ED hashtags[1] (DATE).
- Number of weeks since the member's initial post in the data, measuring the user's *age* (SINCE_START).
- Number of weeks until member's final post in the data (TILL_END).
- The total duration (in weeks) of a member's activity, measuring the user's *lifespan* (DATE_RANGE).
- The appearance (binary) of any variant in the post (VARIANT).
- The appearance (binary) of a variant with a specified edit distance in the post (EDIT_DIST_1, etc.; see § 3.3.2 for a description of how edit distance is computed).
- Maximum orthographic edit distance out of all variants in the post (MAX_EDIT); set to 0 when no variants were in post.

[1] This date is not reported by Instagram but is estimated to be April 1, 2012 (Chancellor et al., 2016).

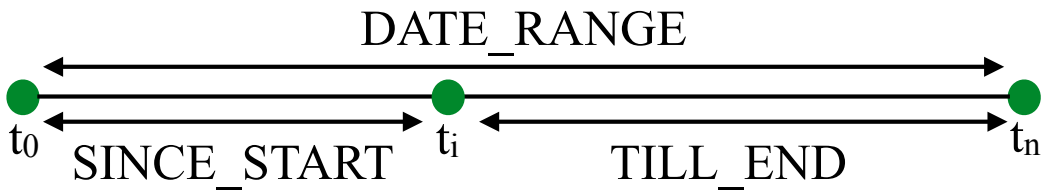


Figure 3.2: Example timeline of member posts at times t_0 (first), t_i and t_n (final) that shows age with statistics SINCE_START and TILL_END, and showing lifespan with DATE_RANGE.

- Total number of all hashtags (variant and non-variant) per post (TAGS).
- Number of comments (COMMENTS) and likes (LIKES) on a post, counted at time of data collection in 2014; log-transformed to adjust for the distributions' long tails.

The distributions of all scalar variables are shown in Figure 3.1. All of the temporal variables have long-tail distributions, indicating that most member lifespans are short.

3.3 Methods

I now outline the methods used in the analysis, including operational definitions for key terms, the edit distance metric used to quantify orthographic variation, and the statistical approaches to address the research questions.

3.3.1 Definitions

For convenience, I provide definitions for the key concepts in the study. I refer to individual Instagram users as "community members", due to their participation in the pro-ED community as signaled by the use of pro-ED hashtags.

- *Age*: for a given post and the associated member, the length of time between the post at time t_i and the first pro-ED post created by the member time t_0. Age is quantified as the variable SINCE_START, which is equal to the number of weeks since the member's first pro-ED post (SINCE_START $= t_i - t_0$). The variable

TILL_END equals the number of weeks until the member's final pro-ED post at time t_n (TILL_END $= t_n - t_i$). These statistics are shown in Figure 3.2. I define a *newcomer* as a member who, at time of posting, has spent less than ten weeks in the community, and a *regular* as a member who, at time of posting, has spent at least ten weeks in the community.[2]

- *Lifespan*: for a given member, the length of time between a member's first and final pro-ED post. Lifespan is quantified as the variable DATE_RANGE, which is equal to the number of weeks between the member's first and final pro-ED post (DATE_RANGE $= t_n - t_0$). This statistic is shown in Figure 3.2. I define a *transient* community member as having a lifespan less than ten weeks in length, and a *committed* member as having a lifespan of at least ten weeks.

- *Source*: any pro-ED hashtag that was banned in April 2012 and has at least one documented orthographic variant; e.g., *#anorexia*.

- *Variant*: any orthographically-varied hashtag that can be associated with a source hashtag; e.g., *#anoreksya*.

- *Depth*: the linguistic distance between a source and its variant: e.g., the variant *#anoreksya* has a depth 3 from its source *#anorexia* (see § 3.3.2).

I acknowledge that the variables SINCE_START, TILL_END, and DATE_RANGE only captured a slice of each community member's behavior, because a pro-ED hashtag member's actual first post on Instagram may be unrelated to pro-ED (and thus unobservable).

[2]It is possible that some members had additional posts in the time between the launch of Instagram in 2010 and the beginning of the data collection in 2011. This would result in an underestimation the age of some individuals. However, the dataset spans four years, and the overwhelming majority of members appear to have ages of less than one year.

3.3.2 Measuring orthographic variation: edit distance

The depth of orthographic variation was quantified by calculating each variant's Levenshtein edit distance from its original form (Levenshtein, 1966). This is equal to minimum number of insertion, deletion and substitution operations necessary to convert a source hashtag to its variant form. For example, transforming *anorexia* to *anoreksya* requires two substitutions ($x \rightarrow k$ and $i \rightarrow y$) and an insertion ($\varnothing \rightarrow s$), thus an edit distance of $1 \times 2 + 1 = 3$. Although in some cases it is useful to design a customized edit distance cost function (Heeringa et al., 2006), in this study all operations were weighted equally for simplicity.[3]

I group orthographic variants by edit distance in Table 3.1 and provide summary statistics for each group, showing the uneven distribution across groups. The variants' frequency grouped by their edit distance is shown in Figure 3.3. Note that the overall frequency of orthographic variants increased over time, particularly for the deeper variants at edit distances 3 and 4. This study examines which community members drove this increase in the frequency and depth of variation over time.

3.3.3 Statistical models

This study used logistic and Poisson regressions as models for their ease of interpretability, since the RQs concern the relative importance of the temporal, social and linguistic variables that may explain variation in hashtag spelling.

I chose a Poisson regression to address the dependent variables (LIKES and COMMENTS in RQ3), because they are count variables with high dispersion and non-normal distributions (Gardner, Mulvey, and Shaw, 1995). The specific regression models for each RQ are described below:

RQ1: Who uses orthographic variants? I used logistic regression to predict

[3]Preliminary tests with a weighted edit distance, with weights derived from the data, showed little difference from the tests with the unweighted edit distance.

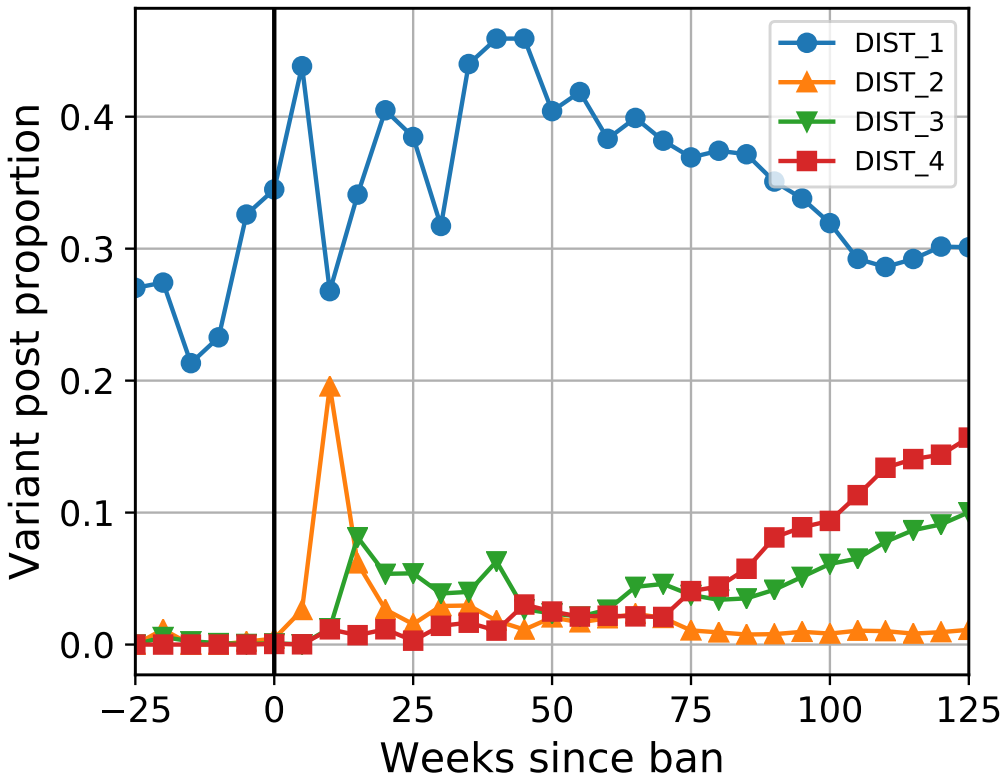


Figure 3.3: Frequency of variants over time, grouped by edit distance: e.g., DIST_1 tracks the normalized frequency of all posts with at least one variant with edit distance 1, such as *#anorexiaa*.

whether a variant spelling appears in a post (dependent variable), using the following membership attributes as independent variables: (1) the post author's relative age (SINCE_START and TILL_END) and (2) the post author's lifespan (DATE_RANGE), as well as the absolute time DATE for both variables.

RQ2: Is depth of variation affected by membership attributes? Depth of variation was measured as the edit distance of a variant from its original form. I considered as a dependent variable the presence of a variant of a specified edit distance from the original tag (e.g., any variant with edit distance 4). I again performed a set of logistic regressions, using the same independent variables as in RQ1 to determine the importance of membership attributes.

RQ3: Does orthographic variation affect social reception? I used Poisson

regressions to predict the number of likes and comments that a post receives (dependent variable), using as independent variables the membership attribute variables of the posting member (DATE_RANGE and SINCE_START) as well as the post's language content (TAGS, MAX_EDIT, VARIANT). I included a fixed effect for each member to account for varying popularity among members.

In all regressions, I removed duplicate posts by members who contribute more than one post for each date to avoid overfitting to the most active members. For logistic regressions, I randomly subsampled the data (N=200,000) and included an equal number of positive and negative labeled instances for class balance. I demonstrate the relative goodness of fit of models using the metric *deviance*, which is a measure of the lack of fit to data (lower values are better). A model's deviance is calculated by comparing the model with the saturated model, defined as the "null model." To interpret the relative importance of the variables in the above regression models, I report the non-standardized coefficients of the regression, p-values (computed through the Wald test, adjusted for Bonferroni correction), and standardized effect sizes (Chinn, 2000). All regressions are performed using the Generalized Linear Model code from the `statsmodels` Python package.[4]

3.4 Results

I addressed the study's RQs by analyzing (§ 3.4.1) the attributes of community members who adopt orthographic variants, (§ 3.4.2) the correlation between orthographic depth and membership attributes, and (§ 3.4.3) the correlation between orthographic depth and social reception. In all regressions, the coefficients β are expressed in terms of the units of the predictors, e.g. log-odds per week. Effect sizes are computed by standardizing the predictors to zero mean and unit variance, and then dividing the resulting coefficients by $\pi/\sqrt{3}$, the standard deviation of the standard logistic distribution (Chinn, 2000).

[4] `http://statsmodels.sourceforge.net/stable/glm.html`

Table 3.2: Regression results for variant appearance in a post, as predicted by relative time variables. *** indicates $p < 0.0001$. In all tables, β indicates the regression coefficient and S.E. indicates the standard error.

Variable	β	S.E.	Effect size
SINCE_START	-4.56E-3***	2.97E-4	-0.348
TILL_END	2.94E-3***	2.88E-4	0.654
DATE	5.29E-3***	1.77E-4	0.746

Table 3.3: Regression results for variant appearance in a post, as predicted by the length of a member's lifespan (observed activity period). *** indicates $p < 0.0001$.

Variable	β	S.E.	Effect size
DATE_RANGE	2.94E-3***	2.89E-4	0.654
DATE	5.41E-3***	1.77E-4	0.746

3.4.1 Orthographic variant authors

The first task is to determine whether a specific author subgroup, such as newcomers (Danescu-Niculescu-Mizil et al., 2013), appeared to drive the community-level tendency toward more orthographic variation. The results of the age regression are displayed in Table 3.2 and the results of the lifespan regression in Table 3.3. The date coefficient is consistently positive across regressions, reflecting a community-level trend toward more variants over real time (see Figure 3.3). I also note consistent coefficients for SINCE_START and TILL_END (negative and positive), revealing a coherent member-level trend away from variants over the member's lifespan in the community. Taken together, these regressions indicate that orthographic variation was perpetuated by newcomers who bring in the new variants and abandon them over the course of their lifespan. The positive coefficient for DATE_RANGE shows that members who will participate or have participated for longer were more likely to use a variant, suggesting that committed members were more prone to participating in the community change.

Both models achieved a better fit than null. The deviance of the null model and the deviance of both models approximately followed a χ^2 distribution, with degrees of

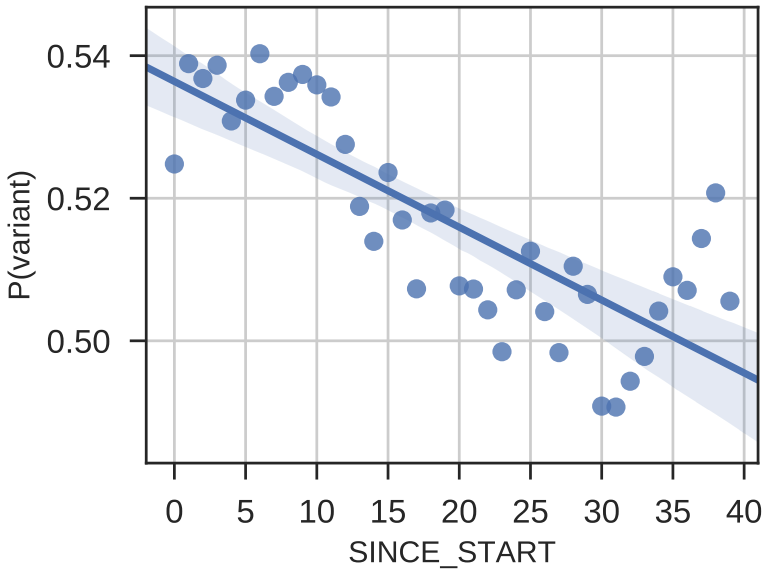


Figure 3.4: Probability of using a variant versus a member's age (weeks since first pro-ED post).

freedom equal to the number of additional variables in the latter model.[5]

This analysis uncovered a split between community-level and member-level variant adoption. As time passed, the overall frequency of orthographic variation increased; but as individual members grew older, they are less likely to use variants, as shown in Figure 3.4. On the other hand, individuals who posted pro-ED content over a long period of time were 4.33% more likely to use a variant than transient community members ($t = 30.9, p < 0.001$). This difference held up for the intersection of the two variables: committed newcomers were 5.09% more likely to use a variant than transient newcomers ($t = 25.4, p < 0.001$).

Overall, *committed* members and *newcomers* were the main contributors to the change toward more frequent orthographic variants.

[5]For age, $\chi^2(3, N = 100,000) = 277,258 - 276,280 = 978$, $p < 10^{-5}$ and for lifespan, $\chi^2(2, N = 100,000) = 277,258 - 276,334 = 924$, $p < 10^{-5}$.

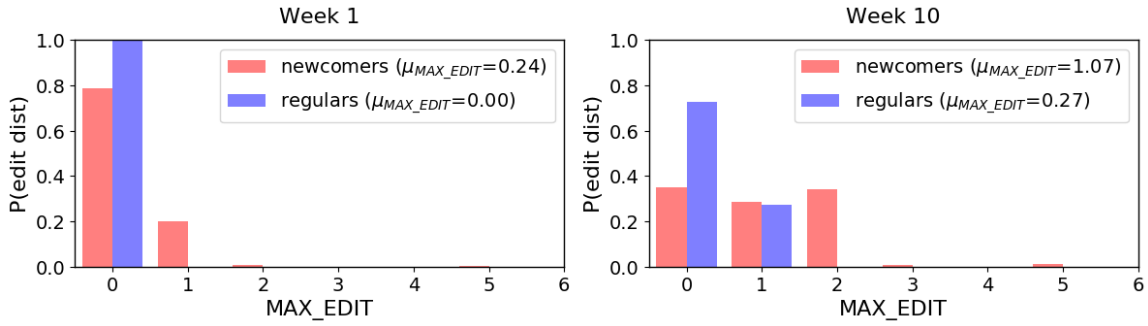


Figure 3.5: Distribution of maximum edit distances across all posts of specified member group (regular versus newcomer) at one week and 10 weeks after the ban (including average edit distance for each group). The newcomers used orthographic variants with consistently higher edit distances than the regulars.

3.4.2 Differences in variant depth by membership

I next examined whether the social correlates of orthographic variation are stronger for variants that are further from the original spelling. The variants were grouped by edit distance, and the strength of association with membership attributes was measured for for low and high edit-distance spellings.

Univariate analysis

Figure 3.5 shows the frequency of posts containing an orthographic variant with edit distances 1-6, broken down by week (since the ban) and by member age (newcomers versus regulars). The newcomers clearly outpaced the regular community members in adopting orthographic variants with higher edit distance.

The impact of member lifespan is shown in Figure 3.6, comparing the average maximum edit distance in posts from committed and transient members of the pro-ED community. Both transient and committed members followed the same community level trend toward using variants with higher edit distance over time, and the separation between transient and committed members remained robust even two years after the ban. To confirm the difference between transient and committed members, I tested all split

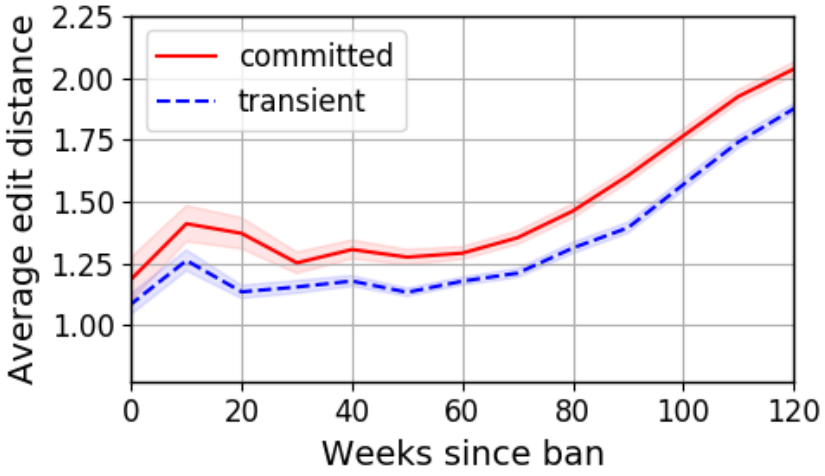


Figure 3.6: Average edit distance over time, binned by DATE and DATE_RANGE and including 95% confidence intervals.

points in the range from 8-12 weeks and found similar results, suggesting that member lifespan can be reliably correlated with orthographic variation.

Multiple logistic regression

To understand the combined impact of member age and lifespan, I use logistic regression, with the outcome variable indicating whether the post contains an orthographic variant of edit distance 1-4 from the source hashtag. The results in Table 3.4 show that effect sizes are larger for the higher edit distance variants, which were more quickly adopted by newcomers, more quickly abandoned by older members, and more strongly favored by committed community members. Social differences therefore correlate not only with the frequency of orthographic variation, but also the depth; conversely, these deeper orthographic variables were better indicators of each member's position in the community.

All edit distance regression models achieve a fit significantly better than the null model: e.g., for the edit distance 4 age regression model, the difference of its deviance from that of the null model approximately follows a χ^2 distribution.[6]

[6] $\chi^2(3, N = 2,416,259) = 277,258 - 253,808 = 23,450, p < 10^{-5}$

Table 3.4: Logistic regression to predict the appearance of a variant with a specified edit distance, as predicted by (1) age and (2) lifespan. *** indicates $p < 0.0001$, * indicates $p < 0.05$.

Model type	Dependent variable: EDIT_DIST_1			Dependent variable: EDIT_DIST_4		
	β	S.E.	Effect size	β	S.E.	Effect size
Age						
SINCE_START	-1.77E-3***	2.98E-4	-0.097	-4.50E-3***	3.12E-4	-0.416
TILL_END	3.11E-3***	2.85E-4	0.250	0.0133***	4.00E-4	1.22
DATE	-1.49E-3***	1.75E-4	-0.127	0.0410***	2.84E-4	3.85
Lifespan						
DATE_RANGE	3.11E-3***	2.88E-4	0.250	0.0133***	4.03E-4	1.22
DATE	-1.49E-3*	1.76E-4	-0.127	0.0344***	2.88E-4	3.85

3.4.3 Social reception of different variants

Finally, I investigated how orthographic variation were received by the community using likes and comments received on a post. Although Chancellor et al. (2016) found that posts with a variant receive more social engagement, it remained to be seen whether this effect is strengthened with deeper edit distance. Since the community norm moves towards variants with deeper edit distance, I expected that posts containing deeper variants would achieve higher engagement in the form of both likes and comments.

To predict the social reception on a given post, I used a Poisson regression, with the outcome variable corresponding to the logarithm of the number of likes and comments for each post.[7] The main predictor was the maximum edit distance of the variants in the post (MAX_EDIT). In addition, I included a number of control predictors: absolute time (DATE), member age (SINCE_START), presence of hashtag variant in post (VARIANT), number of hashtags per post (TAGS), and presence of a source hashtag or one of its variants (e.g., a post with *#ana* and a post with *#anaa* each have a 1 for feature ANA). The hashtag-source variables partly controlled for post topic, since posts about a more popular topic like anorexia might also garner more social reception. Lastly, a fixed effect

[7] I used the *plm* package in R (Croissant and Millo, 2008).

Table 3.5: Poisson regressions for social reception, as predicted by membership and language variables (hashtag coefficients omitted for brevity). *** indicates $p < 0.0001$, otherwise $p > 0.05$. Both models achieve a weak fit: the LOGCOMMENTS regression has R^2=6.82E-3 ($F = 107, p < 0.001$) and the LOGLIKES has R^2=0.0902 ($F = 1550, p < 0.001$).

	Dependent variable: LOGCOMMENTS		Dependent variable: LOGLIKES	
	β	S.E.	β	S.E.
SINCE_START	5.27E-3*	1.57E-3	-0.0319***	9.03E-4
TAGS	0.110***	2.57E-3	0.224***	1.47E-3
VARIANT	-7.89E-3	3.44E-3	-1.14E-3	1.98E-3
MAX_POP	-2.33E-3	1.26E-3	-3.89E-3***	7.25E-4
MAX_EDIT	-3.72E-3	5.51E-3	0.0130***	3.16E-3

for each member was added to control for the possibility that some members receive more social reception than others, due to higher follower counts.

As shown in Table 3.5, posts with deeper variants (higher MAX_EDIT) were positively associated with social engagement through "likes." This complemented the earlier finding that posts with variants received more social attention: increased social attention varied with the depth of variation. However, edit distance was not significantly correlated with comments received, which suggests that posts with especially deep variant hashtags did not elicit the more expensive social signal of a comment, as opposed to the passive "like" signal. This may be due to the relatively high proportion of posts with no comments (heavy left-tail of LOGCOMMENTS in Figure 3.1). As expected, posts with more tags tended to receive more engagement: such posts are easier to find, using Instagram's hashtag-based search functionality. Finally, community members gained fewer likes as they "aged" (negative SINCE_START), possibly because they were actively engaged with other members, or simply because novelty drives interest.

3.5 Limitations and future work

Because Instagram's content ban precluded collecting the data directly (e.g. querying for banned terms), I may have missed some orthographic variants. Furthermore, Instagram's

API did not allow querying for additional member information, such as the date at which each member joined the site instead of the first date at which they used a pro-ED hashtag. This information would have complemented the analysis and helped differentiate newcomers from regulars based on their actual first post date. Having more detailed member information would also provide a better perspective on the correlation between orthographic variation and social reception: for example, testing for a connection between social network structure and orthographic variation.

Future work may explore three possible explanations for the role of newcomers. First, the new pro-ED community community members may have adopted the most extreme practices to signal legitimacy in the community, which represents an extreme version of the community of practice model in which members gain legitimacy through adoption of social and linguistic practices (Lave and Wenger, 1991). Second, the adoption of more extreme hashtag variants may have represented a form of "flag-planting," by which a newcomer attempts to claim a particular hashtag as their own with an especially extreme variant. Third, the supposed newcomer members could actually have been new accounts created as a result of being banned, who then adopted more extreme variants to avoid being banned again. This possibility is especially relevant in the face of prior findings that moderation of deviant behavior online may cause the deviant community member's practices to become more extreme (Cheng, Danescu-Niculescu-Mizil, and Leskovec, 2015). Future work should address the possibility of "repeat offenders" by investigating the relative network of such accounts, as banned accounts may restart their relationships with existing accounts after rejoining with a new username. For this kind of problem, a joint topic-network model could identify accounts who tend to fill the same community role and therefore may be linked accounts, as supported by prior work related to pro-anorexia Twitter posts (Wood, 2015).

3.6 Contributions

- The study extends the notion of variant hashtags to compare individual speaker-level change with community level change. This study found that hashtag spelling can relate to community member status, particularly when the spelling has high relevance to the message content (i.e. whether or not it may be blocked). With respect to RQ1 in the thesis, the consistent patterns among newcomers and long-commitment members revealed that speakers adopt language variants that appear to suit the community even when the community is not well-identified. It is important to relate this result to the fact that newcomers generally adopt innovative language practices, and then retain these practices even as they become outdated with respect to the rest of the community (Danescu-Niculescu-Mizil et al., 2013; Tagliamonte and D'Arcy, 2007). The old-timers in the beer forums studied by Danescu-Niculescu-Mizil et al. (2013) consistently uphold the linguistic habits of their youth. In contrast, the pro-ED Instagram members began with innovative practices, but they abandoned these practices and returned to standard spellings, even as the overall community change was driven by subsequent waves of newcomers toward ever more frequent and deeper orthographic variation. This finding points to a tension in a community member's expected behavior, i.e. avoid censorship and adjust to the actual norms of a community (i.e. lower edit distance over time).

- The study uses orthographic variation to characterize community-level change and differentiating community members by social role. The study provides a new perspective on the community of practice model and shows that overall community norms, even in spelling, can be advanced by newcomers (Danescu-Niculescu-Mizil et al., 2013). For social computing research into community membership, I propose that researchers consider a wider array of possible linguistic norms that are more

situated and less reliant on one-size-fits-all solutions such as LIWC lexicons. For example, future work in social computing should consider whether the behavior of members of a community may be characterized along a continuum, according to the members' linguistic distance from standard language (Tan and Lee, 2015). In a community with relatively standard writing practices, the use of excessive capitalization (e.g., *DUDE*) may be viewed as a less extreme difference from community norms, in comparison with more extreme examples of expressive lengthening (e.g., *duuuuuuudeee*).

CHAPTER 4
CHARACTERIZING COLLECTIVE ATTENTION THROUGH DESCRIPTOR PHRASES

The previous chapter found that members of a hidden community on Instagram consistently adapted to the expected hashtag usage of the community, albeit such that newcomers seemed to overcompensate at first and subsequently reduced their use of hashtag variants. In addition to community standards, people often adjust their language to the expectations of an ongoing event, as they develop common ground with the other event participants and adjust their language to reflect *shared knowledge* expectations (Doyle and Frank, 2015; Galati and Brennan, 2010). For example, repeated exposure to a particular name during an event reduces the need for speakers to explain the importance of the name (Staliūnaitė et al., 2018). While qualitatively understood, the influence of different social factors on the development of knowledge expectations may be hard to assess in natural offline interactions, particularly since it can be hard to study a single person speaking naturally in front of different audiences.

To address this gap, I conclude the section of the thesis related to adjustments to conversation expectations (RQ1) with a study of the public reaction to breaking news events. I find that descriptor use on location names changes regularly in ways that reflect audience expectations: for instance, people tend to drop descriptor information after the peak in post volume during the event. Even in a situation with rapidly changing information and high uncertainty, people consistently add or reduce descriptor information in order to conform to rational communication needs among their unknown audience.

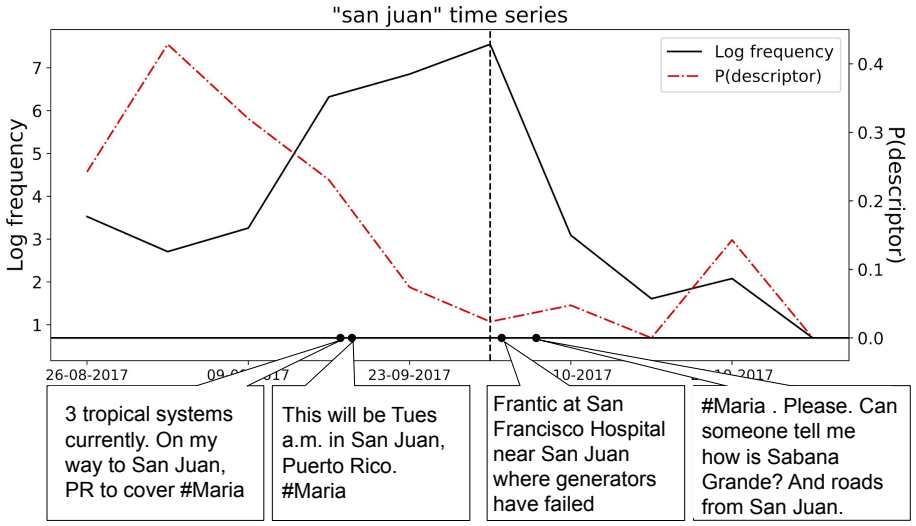
Note: content for this chapter was drawn from Stewart, Yang, and Eisenstein (2020). This work was completed with the help of Diyi Yang and Jacob Eisenstein.

4.1 Motivation

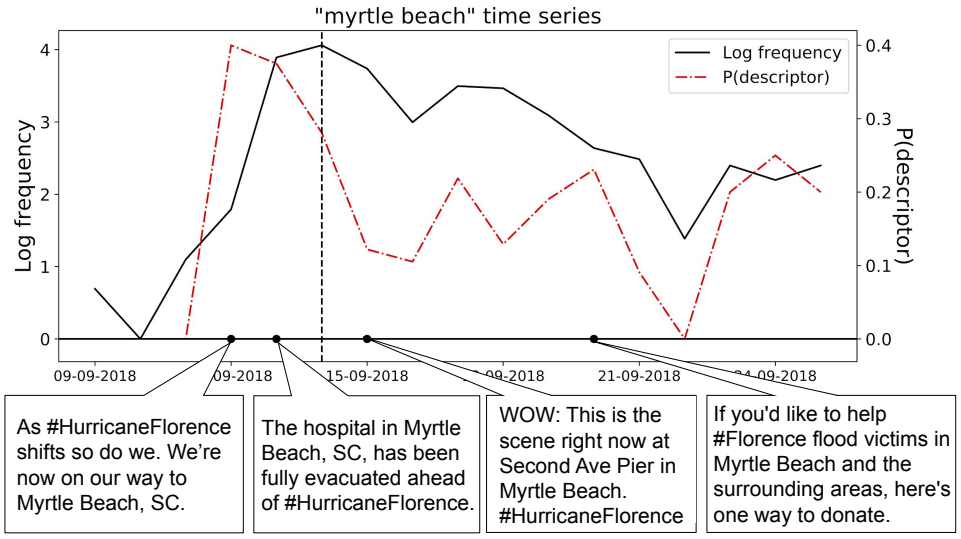
Breaking news events, such as crises, can attract significant *collective attention* from the general public (Lin et al., 2014), resulting in bursts of discussion on social media (Leavitt and Clark, 2014; Lehmann et al., 2012). During such events, attention is often focused on entities (e.g., people, locations, and organizations) that are highly relevant to the unfolding event (Wakamiya et al., 2015). A spike in attention directed toward a particular entity may signal an important update, such as the need for aid for the location (Varga et al., 2013).

While collective attention is often measured with activity metrics such as post volume (Mitra, Wright, and Gilbert, 2016), I characterize collective attention not by *how often* an entity is mentioned, but by *how* it is mentioned. Content-based metrics can be more accurate, as they are less sensitive to data sparsity and biases, such as when a crisis limits internet accessibility. Furthermore, measuring the content of collective attention can provide insight into writer's expectations of reader knowledge, which may become visible as they adapt their writing to address a local or general audience.

To understand how the content of entity references can change with collective attention, consider Hurricane Maria, which struck Puerto Rico in September 2017. As more people became familiar with the locations mentioned in news coverage about the island (DiJulio, Muñana, and Brodie, 2017), news headlines and articles referred to *San Juan* without extra contextual descriptors such as *the capital of Puerto Rico*. This is consistent with a rational model of communication in which linguistic contextualization is used for entities that might otherwise be unknown or ambiguous (Prince, 1992; Staliūnaitė et al., 2018): as San Juan became increasingly salient through repeated mentions, readers could be expected to understand the reference without additional context. Evidence from Twitter in Figure 4.1 supports this hypothesis: the locations of San Juan (Figure 4.1a) and Myrtle Beach (Figure 4.1b) received fewer contextualizing descriptors following peaks in the volume of mentions in discussions of Hurricanes Maria and Florence respectively.



(a) Timeline for mentions of *San Juan* during Hurricane Maria, with example tweets below.



(b) Timeline for mentions of *Myrtle Beach* during Hurricane Florence, with example tweets below.

Figure 4.1: Example of collective attention expressed toward location mentions in discussion of various hurricanes on Twitter. Left y-axis (black solid line) indicates the location's log frequency, right y-axis (red dotted line) indicates the location's probability of receiving a descriptor phrase such as *San Juan, Puerto Rico*. For example, a 25% probability for "*San Juan*" means that 25% of all mentions of *San Juan* had a descriptor phrase.

While global salience plays an important role, more fine-grained factors are also at work. Authors may add or remove context information based on their expectations about their specific audience and on the availability of additional context such as hyperlinks to external articles. Furthermore, even if someone observes an aggregate change in collective attention, one cannot be sure whether the trend is due to change in the author population or a change in the behavior of individual authors. I disentangle these macro-level and micro-level factors in a multivariate analysis of descriptor phrase usage in discussions of crisis events, using data from public discussions of five recent natural disasters on Facebook and Twitter. I investigate the usage of contextual descriptor phrases in references to locations affected by hurricanes, which I link to various proxies for information expectations: temporal trends in relation to the event itself; properties of the author, audience, and entity; and the presence of extra-linguistic context such as hyperlinks and images.

This chapter addresses the following research questions:

- **RQ1**: What factors influence the use of descriptor context in reference to locations of hurricane events?
- **RQ2a**: How does the use of descriptor context for locations change over time at a collective level?
- **RQ2b**: How does the use of descriptor context for locations change at an individual author level?

The main results of the study are as follows. In Facebook posts from public groups concerning Hurricane Maria relief, location mentions received descriptors more often when the locations were not local to the group of discussion, suggesting that descriptors may be used to help explain new information to audiences (§ 4.3.1). In public Twitter posts related to five hurricane events, I find that the aggregate rate of descriptor phrases decreased following the peaks in these locations' collective attention, supporting prior

findings on named entity references in professional newstext (Staliūnaitė et al., 2018) (§ 4.3.2). However, this result is supplemented by more fine-grained effects that also support a rational account of entity reference: authors used fewer descriptors if they had mentioned a location before, and more descriptors if their previous posts received high audience engagement, which suggest a larger (and potentially less contextualized) audience (§ 4.3.3).

This chapter identifies strong connections between entity references and expectations with respect to shared knowledge, which supplements existing linguistic theory while offering researchers and practitioners new tools for measuring and understanding collective attention in crisis events.

4.2 Data

Crisis events present a useful case study for the development of collective attention, due to the large volume of online participation and uncertainty among event observers towards the situation (Varga et al., 2013). This chapter focuses on the collective attention changes in public discourse related to hurricanes, due to hurricanes' lasting economic impact, their broad coverage in the news, and their relevance to specific geographic regions. I collected social media data related to five recent hurricanes. The remainder of this section describes the data collection (§ 4.2.1), location detection (§ 4.2.2), and descriptor detection (§ 4.2.3) for the following datasets:

1. Twitter: 2 million public tweets related to 5 major hurricanes, collected in 2017 and 2018.

2. Facebook: 30,000 posts from 60 public groups related to disaster relief in Hurricane Maria, collected in 2017.

Table 4.1: Summary statistics for Twitter data.

Event	Hashtags	Date range	Tweets	Locations	Location examples
Florence	#florence, #hurricaneflorence	[30-08-18, 26-09-18]	66,595	28,670	*Wilmington, New Bern, Myrtle Beach*
Harvey	#harvey, #hurricaneharvey	[17-08-17, 10-09-17]	679,400	181,636	*Houston, Corpus Christi, Rockport*
Irma	#irma, #hurricaneirma	[29-08-17, 20-09-17]	809,423	229,315	*Miami, Tampa, Naples*
Maria	#maria, #hurricanemaria, #huracanmaria	[15-09-17, 09-10-17]	313,088	57,237	*San Juan, Vieques, Ponce*
Michael	#michael, #hurricanemichael	[06-10-18, 23-10-18]	52,506	22,007	*Panama City, Mexico Beach, Tallahassee*

4.2.1 Collection

Twitter Dataset The Twitter posts were collected using hashtags from five major hurricanes: Hurricane Florence (2018), Hurricane Harvey (2017), Hurricane Irma (2017), Hurricane Maria (2017), and Hurricane Michael (2018). I used hashtags that contained the name of the event in full and shortened form, e.g. #Harvey and #HurricaneHarvey for Hurricane Harvey.

During 2017 and 2018, I streamed tweets that contained hashtags related to the natural disasters at the start of each disaster for up to one week after the dissipation of the hurricane.[1] I augmented this data with additional tweets available in a 1% Twitter sample that contains the related hashtags, restricting the time frame to one day before the formation of the hurricane and one week after the dissipation of the hurricane. Manual inspection revealed minimal noise generated by the inclusion of the name-only hashtags (e.g., #Maria tweets talking about a person named Maria). Summary statistics of the Twitter data are presented in Table 4.1.

I also collected additional event-related tweets from the most frequently-posting authors in each dataset ("active authors"), which were needed to evaluate per-author

[1]Dates are based on estimates from the National Oceanic and Atmospheric Administration (NOAA). For example, estimates for Hurricane Harvey are available here, accessed 1 Jan 2019: `https://www.nhc.noaa.gov/data/tcr/AL092017_Harvey.pdf`.

Table 4.2: Summary statistics for active authors on Twitter.

Event	Authors	Tweets	LOCATION NEs
Florence	186	17,624	29,066
Harvey	164	31,563	50,050
Irma	178	45,913	77114
Maria	139	11,332	18,204
Michael	146	8828	14,655

changes (RQ2b; see § 4.3.3). Table 4.2 summarizes the detailed statistics about the active author data.

Facebook Dataset The Facebook data was collected in the aftermath of Hurricane Maria by searching for public discussion groups that included at least one of Puerto Rico's municipalities in the title (e.g. the group "Guayama: Huracán Maria" refers to Guayama municipality). Relatives and friends of Puerto Ricans often posted in these groups to seek additional information about those still on Puerto Rico, who could not be reached by telephone due to infrastructure damage. I restricted the analysis to Facebook groups related to Hurricane Maria because the limited information available caused more discussion of specific locations, as compared to the other hurricane events that had more up-to-date information available online.

In total, I collected 31,414 public posts from 61 groups, from the time of their creation to one month afterward (September 20 to October 20, 2017). Spanish was the majority language in these posts, so only posts in Spanish were retained, using `langid.py` (Lui and Baldwin, 2012). Due to Facebook data restrictions and API changes, I was unable to collect posts in Facebook groups for the other four hurricane events.

4.2.2 Extracting and filtering locations

I extracted mentions of locations using two systems for named entity recognition (NER): for English, a system that was explicitly adapted to Twitter data (Ritter, Clark, and

Etzioni, 2011)[2] and for Spanish, a general purpose named entity recognizer (Finkel, Grenager, and Manning, 2005).[3] These systems are freely accessible and widely used, and achieve reasonably competitive performance.[4] The performance of these NER systems was evaluted on a sample of tweets (100 tagged LOCATIONs per dataset, 500 total) and found reasonable precision for the `LOCATION` tag (81-96% across all datasets).

For this work, I required named entities that could require descriptor phrases, which include cities and counties. I therefore restricted the analysis to named entities (NEs) that (1) are tagged as `LOCATION`, (2) exist in the GeoNames ontology,[5] (3) map to cities or counties in the ontology, (4) map to *affected* locations in the ontology, based on their location occurring in the region affected by the event, and (5) are unambiguous within the region affected by the event. For instance, the string *San Juan* is a valid location for the Hurricane Maria tweets because the affected region (Puerto Rico) contains an unambiguous match for the string, but it is not a valid location for the Hurricane Harvey tweets because the affected region does not contain an unambiguous match.

4.2.3 Extracting descriptor phrases

One way in which writer can introduce a new entity to a discourse (e.g., *San Juan*) is by linking it to a more well-known entity (e.g., *Puerto Rico*) in a descriptor phrase. To detect this phenomenon, I identified location mentions that had dependent clauses that referred to better-known locations, using population as a proxy. The underlying assumption is that a more well-populated location is be more likely to be known to readers, and can therefore

[2]Accessed 15 Jan 2019: `https://github.com/aritter/twitter_nlp`.

[3]Accessed 15 Jan 2019: `https://nlp.stanford.edu/software/stanford-ner-2018-10-16.zip`.

[4]For location entities, the English tagger has a reported F1 of 74% (Ritter, Clark, and Etzioni, 2011), and the Spanish tagger has a reported F1 of 58% (Finkel and Manning, 2009), but these figures are not directly comparable due to genre differences across datasets. Both English and Spanish are considered "high resource" languages for natural language processing, with hundreds of thousands of tokens of labeled data for named entity recognition (Hovy et al., 2006; Taulé, Martí, and Recasens, 2008). The extension of this data acquisition pipeline to languages that lack substantial labeled data may pose a significant challenge (Rahimi, Li, and Cohn, 2019).

[5]Accessed 15 September 2017: `http://download.geonames.org/export/dump/allCountries.zip`.

Table 4.3: Phrase patterns to capture descriptor phrases in location mentions. Head location marked with underline, context location marked with double underline.

Phrase patterns	Dependency types	Example
LOCATION + LOCATION_STATE	n/a	San Juan, PR
LOCATION + [LOCATION_CONTEXT]$_{\text{MODIFIER}}$	adjective, apposition, preposition, numeric modifier	San Juan, [capital of Puerto Rico]
[LOCATION + LOCATION_CONTEXT]$_{\text{NOUN\_COMPOUND}}$	nominal, compound, apposition	the [Vega Alta neighborhood of San Juan]
LOCATION + [LOCATION_STATE]$_{\text{CONJUNCTION}}$	conjunction	San Juan, Guayama [and Vieques, Puerto Rico]

help describe the preceding location. The frequency of such descriptor phrases is the main dependent variable in this research: I hypothesized that authors are more likely to use such descriptor phrases when they expect readers to treat the location as new information, and less likely to do so when the location is likely to be already known.

To extract sentence structure from text, I used dependency parsing, which decomposes sentences into directed acyclic graphs connecting words and phrases (Eisenstein, 2019). Following Staliūnaitė et al. (2018), a set of dependencies was developed to capture the `MODIFIER` phrase type in a subclause (adjectival clause, appositional modifier, prepositional modifier, numeric modifier) and another set of dependencies to capture the `COMPOUND` type in a super-clause (nominal modifier, compound, appositional modifier). Table 4.3 presents a summary of the phrase patterns that were used to capture descriptor phrases. Taking into account the characteristics of text from two different domains, for the Twitter data I used the `spacy` shift-reduce parser (Honnibal and Johnson, 2015)[6]; for the Facebook data, the dependencies were extracted using the SyntaxNet transition-based parser (Andor et al., 2016).[7] The pilot experiments found that SyntaxNet achieved higher accuracy on Facebook posts, but I was unable to apply it to the larger Twitter dataset due to API restrictions.[8]

[6] `Accessed15Jan2019:https://spacy.io/usage`.

[7] Accessed 15 Jan 2019: `https://cloud.google.com/natural-language/docs/analyzing-syntax`.

[8] As a robustness check I ran the analysis for the Facebook data using parses from `spacy`, and found the same effect directions and significance for all variables considered (see § A.2 in the Appendix).

Table 4.4: Summary of explanatory variables and corresponding metrics, used for descriptor phrase prediction.

Factor	Variable	Description
Importance	Prior location mentions	Frequency of location within the group or event
Author	In-group posts	Posts that an author made within a group
	In-event posts	Posts that an author made about an event (log-transformed)
	In-event posts about location	Posts that an author made about an event that mention the location (log)
	Organization	Whether the author is predicted to be an organization (based on metadata)
	Local	Whether the author is predicted to be local to the event (based on self-reported location)
Audience	Location is local to group	Whether the location exists within the group's associated region
	Group size	Number of unique members who have posted in the group
	Prior engagement	Mean normalized log-count of retweets and likes received by an author (in $t-1$)
	Change in prior engagement	Change in prior engagement received by an author (between $t-2$ and $t-1$)
Information	Has URL	Whether the post contains a URL
	Has image/video	Whether the post contains a URL with an associated image/video
Time	Time since start	Days since first post about event
	During peak	Whether post was written during peak of collective attention toward location
	Post peak	Whether post was written at least 1 day after the peak of collective attention toward location

Validation of extraction performance To assess the accuracy of the phrase patterns in capturing descriptor phrases, two annotators (computer science graduate students) who had not seen the data annotated a random sample of 50 tweets containing at least one location from each data set (250 tweets total). The annotators received instructions on how to determine if a location was marked by a descriptor phrase, including examples that were not drawn from the data, and the annotators marked each location mention as either (1) a "LOCATION + LOCATION_STATE" pattern, (2) one of the other descriptor patterns in Table 4.3 or (3) no descriptor phrase. The annotators achieved high agreement on each separate descriptor type (Cohen's $\kappa = 0.96$ for the state pattern, $\kappa = 0.91$ for the other

patterns). I then filtered posts with perfect agreement, ran dependency parsing on the posts and detected descriptor phrases using the phrase patterns proposed. I found that the phrase patterns achieved reasonable precision and recall (96.6% and 87.5% respectively) in identifying descriptor phrases compared to raters' annotations. This validation check demonstrated that the proposed syntactic patterns can capture descriptor phrases reasonably well.

4.3 Results

I address the research questions in analyses over three separate sets of social factors: static social factors, dynamic factors at the collective level, and dynamic factors at the individual level.

4.3.1 Non-temporal social factors in descriptor use

I first address RQ1, concerning which social factors influence the use of descriptor context when referring to locations of hurricane events. Of particular interest are the indicators of whether locations may be considered shared knowledge within a community. A descriptor phrase may be omitted for locations that are geographically local to a group of people, i.e. knowledge already shared among the group (e.g., if someone mentions the location *San Juan* in a group based in a region containing San Juan).

I compared the rate of descriptor uses for location mentions in both Facebook and Twitter. For the Facebook data, I determined whether the group's region contains the location mentioned based on whether the most likely match for the location in the gazetteer is contained in that region.[9] I also considered the following additional predictors: frequency with which the location is mentioned in prior posts (importance), author posting frequency in the group (author status), and group size (audience), as summarized in Table 4.4. For the Twitter data, I considered the following predictors:

[9]When a location string matches multiple location entities, I choose the one with the highest population.

location mention frequency in the Twitter sample (importance), whether the author is an organization or a local to the location (author status),[10] whether the post has a URL (information), and whether the post has an image or video (information).

I used separate logistic regression models for the Facebook and Twitter data. In both cases, the dependent variable was whether each location mention was accompanied by a descriptor phrase (N=18,432 and N=49,020, respectively). In detail, I used an elastic net regression (Zou and Hastie, 2005) in order to reduce the risk of overfitting.[11] For this analysis, rare categorical values ($N < 20$) for the fixed effects were replaced with the default `RARE` value to avoid overfitting to uncommon categories. The columns "RQ1 (Facebook)" and "RQ1 (Twitter)' in Table 4.5 report the results of the logistic regression.

On Facebook (see "RQ 1 (Facebook)" in Table 4.5), mentions of locations that were local to the group received significantly fewer contextual descriptors ($\beta = -0.623, p < 0.001$). In the group "Hurricane Maria in Lajas" the mention of the municipality *Lajas* did not receive an descriptor (*Do you know if the Bank is open in Lajas?*), while in the group "Guayama: Huracán Maria" the mention of *Lajas* did receive an descriptor (*People who can bring water to Lajas Puerto Rico: they need water urgently*).[12] The other predictors did not have a statistically significant effect on descriptor use.

On Twitter, there were several significant effects: more salient and important locations received fewer descriptors ($\beta = -0.172, p < 0.001$); authors who were local to an event were less likely to include descriptor phrases ($\beta = -0.511, p < 0.001$), while organizational accounts were more likely to use descriptor phrases ($\beta = 0.093, p < 0.01$); in posts that contain URLs, descriptors were less likely to appear ($\beta = -0.081, p < 0.05$), but in posts that linked to an image or a video, descriptors were more likely ($\beta = 0.137, p < 0.001$).

[10] See § A.1 for details on determining whether an author is an organization or local.

[11] An L2 regularization of 0.01 was chosen through grid search to maximize log-likelihood on held-out data (90-10 train/test split).

[12] Comments are translated from Spanish and paraphrased to preserve privacy.

Table 4.5: Logistic regression results for all analysis, predicting the presence of a descriptor phrase. * indicates $p < 0.05$, otherwise $p > 0.05$.

		RQ1 (Facebook)		**RQ1** (Twitter)		**RQ2a** (Twitter)		**RQ2b** (Twitter)	
Factor	Variable	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Intercept		-2.030	28.550	-1.052*	0.404	-1.026*	0.415	-1.222	11.206
Importance	Prior location mentions	-0.075	7.164	-0.172*	0.025	-0.200*	0.031	-0.107	0.114
Author	Author in-group posts	-0.328	0.522	-	-	-	-	-	-
	Author is organization	-	-	0.093*	0.033	0.092*	0.035	-0.149	0.115
	Author is local	-	-	-0.511*	0.020	-0.797*	0.031	-0.671*	0.107
	Prior event-based posts (from author)	-	-	-	-	-	-	0.110	0.093
	Prior location mentions (from author)	-	-	-	-	-	-	-0.237*	0.091
Audience	Local location	-0.623*	0.106	-	-	-	-	-	-
	Group size	0.121	0.040	-	-	-	-	-	-
	Prior engagement (author)	-	-	-	-	-	-	0.292*	0.052
	Change in prior engagement (author)	-	-	-	-	-	-	-0.004	0.042
Information	Has URL	-	-	-0.081*	0.035	-0.058	0.038	-0.482*	0.154
	Has image/video	-	-	0.137*	0.032	0.124*	0.034	0.562*	0.123
Time	Time since start	-	-	-	-	-0.120*	0.036	-0.004	3.63
	During-peak	-	-	-	-	0.004	0.038	0.144	0.122
	Post-peak	-	-	-	-	-0.127*	0.049	-0.189	0.157
Model deviance			469		2954		4127		2239
Accuracy			71.3%		72.7%		73.3%		75.0%

These findings match with the view that authors customize their information sharing based on the perceived needs of readers. Additional context was unnecessary when writing for locals, or when writing about entities that are already salient, or have become salient through repeated mentions. Twitter accounts that represented large organizations were likely expecting larger audiences who require more context; locals were more likely writing for their peers, who do not require context. Additional context can be provided by hyperlinks to detailed stories, but multimedia content such as images and videos does not serve the same purpose, and therefore required additional contextualization for an audience who may not immediately understand the importance of the images.

4.3.2 Collective change in descriptor context use

I next turn to a temporal analysis of descriptor use, using longitudinal data from Twitter. As collective attention focuses on affected locations over the course of a crisis event, I expect those locations to require less contextualization. To test this theory, I augmented the predictors from the previous section with two temporal variables: whether the message is posted during or after the peak volume in the discussion of the event, and how many days have elapsed since the start of the hurricane.

The definition of the peak in collective attention is critical, because it determines the point at which an entity is expected to become shared knowledge in a discussion (Staliūnaitė et al., 2018). Following Mitra, Wright, and Gilbert (2016), I defined the time of peak collective attention $\hat{t}_i$ for each location i as the 24-hour period during which it is mentioned the most frequently:

$$\hat{t}_i = \arg\max_{t \in T} f_t^{(i)}$$

where $f_t^{(i)}$ is the raw frequency of location i at time t (see Figure 4.1 for peak in *San Juan* and *Myrtle Beach* mentions). I defined *pre-peak* as the period that ends t_{buffer} days before

the frequency peak, *during-peak* as the period at most t_{buffer} days before and at most t_{buffer} days after, and *post-peak* as the period that begins t_{buffer} days after the frequency peak (setting $t_{\text{buffer}} = 1$). As described below, fixed effects for authors and locations were included in robustness checks; to improve stability, I removed all locations that are mentioned on fewer than $N = 5$ separate dates, and combined all authors with only a single post into a `RARE` bin.

As shown in the "RQ2a (Twitter)" column of Table 4.5, the post-peak time period had less descriptor use than the earlier time periods ($\beta = -0.127, p < 0.001$). Furthermore, descriptor phrase use decreased with the number of days since the start of the event ($\beta = -0.120, p < 0.001$). These findings are consistent with the hypothesis that entities become more salient through the focus of collective attention, and that this salience makes contextualization less necessary. The regression also provides more rigorous validation for the trend shown in Figure 4.1.

However, an additional potential explanation for the decrease in descriptor context may be a change in the set of authors after the peak in collective attention – for example, an influx of locals, who are less likely to use descriptors overall. To test for this, I re-ran the regression above and replaced the author variables ("local" and "organization") with a fixed effect for each author. Here, the post-peak effect was still significant and negative ($\beta = -0.253, p < 0.05$). This suggests that a change in author population did not explain the decrease in descriptor use over time, or else this would have been absorbed by the fixed effects. I note that these findings generally replicated prior work on long-term trends in descriptor phrase usage in non-crisis contexts (Staliūnaitė et al., 2018), although prior work did not consider attention peak as the time variable.

4.3.3 Individual change in descriptor context use

I now further examine temporal dynamics at the level of individual authors (RQ2b). Under a strong interpretation of the motivating hypotheses, an author who participated frequently

in early discussion of the event would use fewer descriptors later during the event, under the assumption that their readers would no longer need context for their event-related posts. However, other factors may also be at work: an author who had a growing audience may be more likely to use descriptor phrases to accommodate their new readers.

To better model the author-level changes in descriptor use, I introduced the following additional predictors: number of prior posts by author during event (author-level), number of prior posts by author about the location during event (author-level), engagement received by author at $t-1$ (audience),[13] and change in engagement received by author between $t-2$ and $t-1$ (audience). These predictors required a longitudinal sample of frequently-posting authors, i.e. *active* authors, who were identified as those whose post volumes were at or above the 95^{th} percentile among all authors in the collection. I scraped all publicly available tweets posted by these active authors that mention one of the event's hashtags during the event time period (e.g., all posts for a Harvey-related active author from between August 17 and September 10, 2017 that use #Harvey or #HurricaneHarvey). The locations and descriptor phrases were processed as described in § 4.2.2, and I report the relevant statistics for these active authors in Table 4.2. I used similar regularized logistic regression models with only data from the active authors who posted at least once during each of the time periods, so as to isolate changes for individual authors.

The results are described in the "RQ2b (Twitter)" column of Table 4.5. I found that authors' prior mentions of a location were associated with less descriptor use ($\beta = -0.237, p < 0.001$) but that there was no significant temporal trend with respect to the start of the event or the peak attention. This latter null result held even when I performed the regression without the additional *author* and *audience* variables. I did find that authors who received more engagement from the audience tended to use more descriptors ($\beta = 0.292, p < 0.001$), which is again consistent with the view that larger audiences necessitate additional contextualization.

[13] I define engagement as the mean of retweets and likes, converted into z-scores across the population.

I hypothesized that the active authors may be different from the overall population in how they respond to trends in collective attention. To test this, I identified "less active" authors as those with lower post volumes below the 95^{th} percentile, and I re-ran the regression analysis with only these individuals, using the same variables as the active authors. I found that these less active authors did show a significant decrease in descriptor use following the peak in collective attention ($\beta = -0.127, p < 0.05$) and a decrease in descriptor use over time ($\beta = -0.098, p < 0.05$), suggesting a qualitative difference between the less active and more active authors.

It is unclear whether highly active authors had special characteristics, or whether these differences were driven by some other aspect of the design. The set of active authors contains many journalists and news outlets, whose patterns of writing may be shaped by stylistic formalisms but also a greater sensitivity to their audience's awareness of unfolding situations (Murthy and Gross, 2017). One verified news account covering Hurricane Irma tended to include descriptors for names like *Jacksonville* early on (10 Sept 2017: *"No access to hospitals" once winds reach a sustained 30 mph near Jacksonville, Florida*) and later dropped descriptors after mentioning these names repeatedly (12 Sept 2017: *Jacksonville sheriff's office hopes the people they rescued "will take evacuation orders seriously"*).[14] This suggests that the news account assumed their audience would be following their earlier mentions of *Jacksonville* and would adapt to the lack of context.

4.4 Limitations and future work

This study focused on a small set of specific crisis events, chosen mainly due to the large volume of online discussions and high uncertainty of the unfolding events. It is possible that the patterns observed in the study were specific to these events and locations, so more work is required to establish generalization to other types of crisis events and other types

[14] Content paraphrased to preserve privacy.

of entities. Even within these events, the data collection relied on hashtags that may not capture the full breadth of discussion of the crisis events, because I may have missed less frequent hashtags that covered other aspects of the discussion. If these hashtags were for some reason unrepresentative, then a more extensive dataset might reveal other relationships between descriptor use and information expectations. This study focused exclusively on location names because of their geographic relevance to events, but future work should examine other types of named entities (people, organizations) that also undergo change in response to increased attention (Staliūnaitė et al., 2018). Finally, I acknowledge that other online contexts might give rise to different expectations about audience knowledge, e.g., fast-paced forums for news readers versus online encyclopedias whose text is meant to be relevant long after the crisis has passed. Even on the same platform, different audience contexts may yield different patterns of conversation adaptation, e.g. @-replies among friends could indicate more common ground (Doyle and Frank, 2015) and therefore less need for descriptors. The specific social context used in this study may capture idiosyncratic patterns in the use of descriptors for contextualizing information.

Future work should investigate more long-term examples of descriptor use change in news media, including cases where descriptors may reemerge after being dropped as in the case of Flint, Michigan (often referred to as *Flint* in the early stages of the water crisis that started in 2014). Identifying common trajectories of descriptor use and writing styles across events can provide insight into how public information needs may shift in response to changing crisis conditions (Olteanu, Vieweg, and Castillo, 2015). With respect to crises, follow-up work should investigate different types of crisis events to determine whether expectations of shared knowledge are significantly different based on the circumstances (Houston et al., 2015). A fast-moving and highly lethal crisis such as an earthquake may require news media to drop context information quickly to make way for newer or more important information, while a more slow-moving crisis such as a

pandemic may allow media to retain context to accommodate new readers.

From the linguistic perspective, future work should consider alternate definitions of "context" beyond the descriptor phrases used in this study. Longer spans of text may include context information in less direct ways (e.g. *San Juan is flooding. It is the capital of Puerto Rico*) that may still reflect the author's assumed need for context. Even in the narrow context of descriptor phrases, more granular definitions of descriptors could reveal fine-grained trends, e.g. whether authors use less well-known "anchors" in descriptors as in *Ichikawa, in Tokyo* versus *Ichikawa, in the Chiba prefecture*, since *Chiba* would be more useful for local speakers. Considering a more granular definition of description context can also help adapt this kind of analysis to events that may be more narrow than hurricanes (e.g. protests; Gleason, 2013), which would likely require more fine-grained references among participants to help navigate uncertainty (e.g. street names, local points of interest).

4.5 Contributions

- With respect to the first research question of the thesis related to conversation expectations, this study reveals three major trends in descriptor use that correspond to shared knowledge expectations:

    1. When authors were local to a place, or wrote for an audience who was expected to be local, they were less likely to use descriptor phrases to contextualize references to locations, reflecting shared knowledge among the author and audience.
    2. At a collective level, authors used fewer descriptors over the course of crisis events even after controlling for a set of explanatory variables.
    3. At an individual level, highly active authors changed their descriptor use in response to prior audience engagement but not after the peak in collective

attention, whereas relatively less active users showed a significant decrease in such context use over time.

Taken together, these trends in descriptor use suggest a consistent awareness of audience information expectations to which authors readily adapt. Even in a difficult situation such as a crisis, social media authors actively shape their writing to achieve their communication goal of making sense of an uncertain situation.

The overall findings supports the notion of rational communication (Grice, 1975) by which speakers seek to contribute the most relevant information to a given discussion. This study shows that aggregate-level responses to news events can be understood as a collective sense-making process (Heverin and Zach, 2012) by which individual people work to improve their individual understanding of an uncertain situation. Far from an inevitable monotonic process toward fewer descriptors over time (Staliūnaitė et al., 2018), the use and non-use of extra linguistic structure fluctuates in accordance with situational communication needs, such as whether the author is local to the event or highly active.

- The study provides a useful framework for social computing researchers to quantify collective attention expectations with descriptive information. The metric involved is lightweight, requiring only named entity recognition and short-distance parsing, and can be modified to address different notions of syntactic context. Furthermore, the metric is robust to changes in the data sampling rate, since the use or non-use of a descriptor is a proportion that can be measured accurately with any amount of data and does not need to be normalized (unlike e.g. word frequency).

  The same framework can be extended to other instances of large-scale discussions that exhibit uncertainty among participants, such as reactions to protests and political upheaval (Garimella et al., 2017; Heverin and Zach, 2012). The metric may even provide insight to crisis organizations that need a view of the public's

understanding of an unfolding situation, when such information may not be available from official data. For instance, a crisis organization may investigate the descriptive context of a highly-discussed location on social media to understand whether the public has prior information about the location, which may then help the organization decide what information the public needs.

4.6 Thesis section summary

This study concludes the section of the thesis that examines RQ1, i.e. how speakers adjust their language to the assumed expectations of their community and their discussions when the other participants are not well-defined. I find that people leveraged hashtag spelling to adjust to the assumed expectations of their community, and people also adjusted their use of descriptor phrases to dynamically meet the rational expectations of their audience in breaking news events. Taken together, these studies demonstrate that speakers change their behavior to fit the norms of their respective conversations without necessarily having a strong social bond to other participants, which may relate to their intention to share their experiences for the benefit of the broader community or discussion. These two studies also reveal that speakers' adapting to conversation expectations can be partly shaped by the affordances of social media, such as the use of hashtags to shape the potential future audience and the role of social engagement as an incentive for behavior change.

I now move to language change as the next topic of interest, concerning the long-standing sociolinguistic inquiry into the relative influence of different global factors on language change. Social media provides a broad view of societal changes that can happen rapidly, most notably the change in language norms. The next thesis chapter investigates large-scale language change in online discussions, to address the open sociolinguistic question of how different structural factors contribute to the adoption of new words.

CHAPTER 5
WORD GROWTH AND DECLINE

The previous studies considered the degree to which people adapt to the assumed expectations of other conversation participants when joining an online community and joining discussion of ongoing news events. I now consider a broader question in with respect to language change (RQ2): how readily do linguistic context dissemination and social context dissemination explain the adoption of words in online communities? This connects to the broader question of innovation dissemination, i.e. how new norms are spread across the internet (Goel et al., 2016; Leskovec, Backstrom, and Kleinberg, 2009; Tsur and Rappoport, 2015). While studies in innovation spread often focus on the *social* side of innovations (e.g. who are the people leading the change?), such studies tend to ignore the fact that new words are bound to the structure of the language in which they originate (Kershaw, Rowe, and Stacey, 2016; Ryskina et al., 2020). For instance, are words that are more linguistically flexible also seen as more useful by speakers and therefore more likely to be adopted? Rather than treating words as atomic units to be adopted or abandoned, studies of language change must consider the structural context in which the words can be used, in order to better model the words' inherent utility.

The next chapter studies the growth and decline of words in online communities. Specifically, I analyze long-term word frequency change on Reddit with the goal of disentangling the relative influence of different social and linguistic factors. I identify nonstandard words that exhibit consistent growth and decline, then I leverage several metrics to quantify their dissemination across social and linguistic contexts. I find that linguistic dissemination proves more important than social dissemination in explaining word growth and decline overall, which suggests a limit to the power of social structures in explaining language change.

Note: content for this chapter was drawn from Stewart and Eisenstein (2018). This work was completed with the help of Jacob Eisenstein.

5.1 Motivation

With the fast-paced and ephemeral nature of online discourse, language change in online writing is prevalent (Androutsopoulos, 2011) and noticeable (Squires, 2010). In social media, new words emerge constantly to replace even basic expressions such as laughter: today's *haha* is tomorrow's *lol* (Tagliamonte and Denis, 2008). Why do some nonstandard words, like *lol*, succeed and spread to new contexts, while others, like *fetch*, fail to catch on? Can a word's growth be predicted from patterns of usage during its early days?

Language change can be treated like other social innovations, such as the spread of hyperlinks (Bakshy et al., 2011) or hashtags (Romero, Meeder, and Kleinberg, 2011; Tsur and Rappoport, 2015). A key aspect of the adoption of a new practice is its *dissemination*: is it used by many people, and in many social contexts? High dissemination enables words to achieve greater exposure among social groups (Altmann, Pierrehumbert, and Motter, 2011), and may signal that the innovation is positively evaluated.

In addition to social constraints, language change is also shaped by grammatical constraints (D'Arcy and Tagliamonte, 2015). New words and phrases rarely change the rules of the game but must instead find their place in a competitive ecosystem with finely-differentiated linguistic roles, or "niches" (MacWhinney, 1989). Some words become valid in a broad range of linguistic contexts, while others remain bound to a small number of fixed expressions. I therefore posit a structural analogue to social dissemination, which I call *linguistic dissemination*.

I compare the fates of such words to determine how linguistic and social dissemination each relate to word growth, focusing on the adoption of nonstandard words in the popular online community Reddit. The following hypotheses are evaluated:

- **H1: Nonstandard words with higher initial social dissemination are more likely**

to grow. Following the intuition that words require a positive social evaluation to succeed, I hypothesize a positive correlation between social dissemination and word growth.

- **H2-weak: Nonstandard words with higher linguistic dissemination in the early phase of their history are more likely to grow.** This follows from work in corpus linguistics showing that words and grammatical patterns with a higher diversity of collocations are more likely to be adopted (Ito and Tagliamonte, 2003; Partington, 1993).
- **H2-strong: Nonstandard words with higher linguistic dissemination are more likely to grow, even after controlling for social dissemination.** This follows from the intuition that linguistic context and social context contribute differently to word growth, and that a word's inherent utility presents benefits that are separate from a word's social capital.

To address H2, I develop a novel metric for characterizing linguistic dissemination (§ 5.3.2), by comparing the observed number of n-gram contexts to the number of contexts that would be predicted based on frequency alone. I then analyze the relative effect of social and linguistic dissemination using the following analysis methods: prediction of frequency change in growth words (as in prior work) (§ 5.4.1); causal inference of the influence of dissemination on probability of word growth (§ 5.4.2); binary prediction of future growth versus decline (§ 5.4.3); and survival analysis, to determine the factors that predict when a word's popularity begins to decline (§ 5.4.4).

All tests indicate that linguistic dissemination plays an important role in explaining the growth and decline of nonstandard words, more so than social dissemination.

5.2 Data

The study examines the adoption of words on social media, and I focus on Reddit as a source of language change. Reddit is a social content sharing site separated into distinct

Table 5.1: Data summary statistics.

	Total	Monthly mean
Comments	1,625,271,269	45,146,424
Tokens	56,674,728,199	1,574,298,006
Subreddits	333,874	48,786
Users	14,556,010	2,302,812
Threads	102,908,726	3,079,780

sub-communities or "subreddits" that center around particular topics (Gilbert, 2013). Reddit is a socially diverse and dynamic online platform, making it an ideal environment for research on language change (Kershaw, Rowe, and Stacey, 2016; Tredici and Fernández, 2018). Furthermore, because Reddit data is publicly available I expect that this study can be more readily replicated than a similar study on other platforms such as Facebook or Twitter, whose data is less easily obtained.

I analyze a set of public monthly Reddit comments[1] posted between 1 June 2013 and 31 May 2016, totalling $T = 36$ months of data. This dataset has been analyzed in prior work (Hessel, Tan, and Lee, 2016; Tan and Lee, 2015) and has been noted to have some missing data (Gaffney and Matias, 2018), although this issue should not affect the analysis, assuming a random distribution of missing data. To reduce noise in the data, I filtered all comments generated by known bots and spam users[2] and filtered all comments created in well-known non-English subreddits.[3] The final data collected is summarized in Table 5.1.

I replaced all references to subreddits and users (marked by the convention *r/subreddit* and *u/user*) with *r/SUB* and *u/USER* tokens, and all hyperlinks with a *URL* token. I also reduced all repeated character sequences to a maximum length of three (e.g., *loooool* to *loool*). The final vocabulary includes the top 100,000 words by frequency.[4] All

[1]Accessed 1 Oct 2016: `http://files.pushshift.io/reddit/comments/`.

[2]The same list used in Tan and Lee (2015), accessed 1 October 2016: `https://chenhaot.com/data/multi-community/README.txt`.

[3]I randomly sampled 100 posts from the top 500 subreddits and labelled a subreddit as non-English if fewer than 90% of its posts were identified by `langid.py` (Lui and Baldwin, 2012) as English.

[4]I restricted the vocabulary because of the qualitative analysis required to identify nonstandard words.

OOV words were replaced with UNK tokens, which comprise 3.95% of the total tokens.

5.2.1 Finding growth words

The work seeks to study the growth of nonstandard words, which I identify manually instead of relying on pre-determined lists (Tredici and Fernández, 2018). To detect such words, I first computed the Spearman correlation coefficient between the time steps $\{1...T\}$ and each word w's frequency time series $f^{(w)}_{(1:T)}$ (frequency normalized and log-transformed). The Spearman correlation coefficient captures monotonic, gradual growth that characterizes the adoption of nonstandard words (Grieve, Nini, and Guo, 2016; Kershaw, Rowe, and Stacey, 2016).

The first set of words was filtered by a Spearman correlation coefficient above the 85^{th} percentile ($N = 15,017$). From this set of words, 1,120 words in set $\mathcal{G}$ ("growth") were identified that were neither proper nouns (*berniebot, killary, drumpf*) nor standard words (*election, voting*). These words were removed because their growth may be due to exogenous influence. A standard word is one that can plausibly be found in a newspaper article, which follows from the common understanding of newspaper text as a more formal and standard register. Therefore, a nonstandard word is one that cannot plausibly be found in a newspaper article, a judgment often used by linguists to determine what counts as slang (Dumas and Lighter, 1978). In ambiguous cases, a sample of comments that included the word were inspected to determine consistency of meaning. A colleague and I annotated the top 200 growth candidates for standard/proper versus nonstandard (binary), which yielded sufficiently high inter-annotator agreement of κ=0.79.

5.2.2 Finding decline words

To determine what makes the growth words successful, the study required a control group of "decline" words, which are briefly adopted and later abandoned. Although these words may have been successful before the time period investigated, their decline phase makes

them a useful contemporaneous comparison for the growth words. I found such words by fitting two parametric models to the frequency series of all words.

Piecewise linear fit I fit a two-phase piecewise linear regression on each word's frequency time series $f_{(1:T)}$, which splits the time series into $f_{(1:\hat{t})}$ and $f_{(\hat{t}+1:T)}$. The goal was to select a split point $\hat{t}$ to minimize the sum of the squared error between observed frequency f and predicted frequency $\hat{f}$:

$$\hat{f}(m_1, m_2, b, t) = \begin{cases} b + m_1 t & t \leq \hat{t} \\ b + m_1 \hat{t} + m_2(t - \hat{t}) & t > \hat{t}, \end{cases} \tag{5.1}$$

where b is the intercept, m_1 is the slope of the first phase, and m_2 is the slope of the second phase. Decline words $\mathcal{D}_p$ ("piecewise decline") displayed growth in the first phase ($m_1 > 0$), decline in the second phase ($m_2 < 0$), and a strong fit between observed and predicted data, indicated by $R^2(f, \hat{f})$ above the 85^{th} percentile (36.1%); this filtering yielded 14,995 candidates.

Logistic fit To account for smoother growth-decline trajectories, I also fit the growth curve to a logistic distribution, which is a continuous unimodal distribution with support over the non-negative reals. I identified the set of candidates $\mathcal{D}_l$ ("logistic decline") as words with a strong fit to this distribution, as indicated by R^2 above the 99^{th} percentile (82.4%), yielding 998 candidates. The logistic word set partially overlapped with the piecewise set, because some words' frequency time series show a strong fit to both the piecewise function and the logistic distribution.

Combined set I combined the sets $\mathcal{D}_p$ and $\mathcal{D}_l$ to produce a set of decline word candidates ($N = 15,665$). Next, I filtered this combined set to exclude standard words and proper nouns, yielding a total of 530 decline words in set $\mathcal{D}$. Each word was assigned a split point $\hat{t}$ based on the estimated time of switch between the growth phase and the

Table 5.2: Examples of nonstandard words in all word sets: growth ($\mathcal{G}$), logistic decline ($\mathcal{D}_l$) and piecewise decline ($\mathcal{D}_p$).

Word set	Examples
$\mathcal{G}$	*idk, lmao, shitpost, tbh, tho*
$\mathcal{D}_l$	*atty, eyebleach, iifym, obeasts, trashy*
$\mathcal{D}_p$	*brojob, nparent, rekd, terpers, wot*

Table 5.3: Word formation category counts in growth ($\mathcal{G}$) and decline ($\mathcal{D}$) word sets.

	Clipping	Compound	Respelling	Other	Total
$\mathcal{G}$	198 (17.7%)	334 (29.8%)	83 (7.4%)	505 (45.1%)	1,120
$\mathcal{D}$	53 (10.0%)	100 (18.9%)	108 (20.4%)	269 (50.8%)	530

decline phase, which was the split point $\hat{t}$ for piecewise decline words and the center of the logistic distribution $\hat{\mu}$ for the logistic decline words.

Examples of both growth and decline words are shown in Table 5.2. The growth words include several acronyms (*tbh*, "to be honest"; *lmao*, "laughing my ass off"), while the decline words include clippings (*atty*, "atomizer"), respellings (*rekd*, "wrecked"; *wot*, "what") and compounds (*nparent*, "narcissistic parent").

I also provide a distribution of the words across word generation categories in Table 5.3, including compounds and clippings in similar proportions to prior work (Kulkarni and Wang, 2018). Because the growth and decline words exhibited similar proportions of category counts, the word category did not present a significant confound in differentiating growth from decline.

5.3 Methods

I now outline the methods used to compute the degree of **social** and **linguistic** dissemination in the growth and decline words.

5.3.1 Social dissemination

I rely on the dissemination metric proposed by Altmann, Pierrehumbert, and Motter (2011) to measure the degree to which a word occupies a specific social niche (e.g., low dissemination implies limited niche). To compute user dissemination D^U for word w at time t, I first computed the number of individual users who used word w at time t, written $U_t^{(w)}$. I then compared this with the expectation $\tilde{U}_t^{(w)}$ under a model in which word frequency is identical across all users. The user dissemination is the log ratio,

$$\log \frac{U_t^{(w)}}{\tilde{U}_t^{(w)}} = \log U_t^{(w)} - \log \tilde{U}_t^{(w)}. \tag{5.2}$$

Following Altmann, Pierrehumbert, and Motter (2011), the expected count $\tilde{U}_t^{(w)}$ was computed as,

$$\tilde{U}_t^{(w)} = \sum_{u \in \mathcal{U}_t} (1 - e^{-f_t^{(w)} m_t^{(u)}}), \tag{5.3}$$

where $m_t^{(u)}$ equals the total number of words contributed by user u in month t, and $\mathcal{U}_t$ is the set of all users active in month t. This corresponds to a model in which each token from a user has identical likelihood $f_t^{(w)}$ of being word w. In this way, I computed dissemination for all users (D^U), subreddits (D^S) and threads (D^T) for each month $t \in \{1...T\}$.

5.3.2 Linguistic dissemination

Linguistic dissemination captures the diversity of linguistic contexts in which a word appears, as measured by unique n-gram counts. I computed the log count of unique trigram[5] contexts for all words (C^3) using all possible trigram positions: in the sentence *that's cool af haha*, the term *af* appears in three unique trigrams, *that's cool af, cool af haha, af haha <END>*.

[5]Pilot analysis with bigram contexts gave similar results.

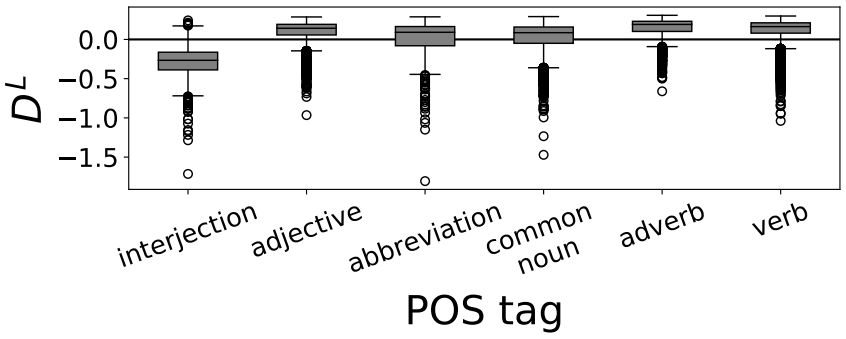


Figure 5.1: Distribution of mean linguistic dissemination (D^L) across part of speech groups.

The unique log number of trigram contexts was strongly correlated with log word frequency ($\rho(C^3, f) = 0.904$), as implied by Heaps' law (Egghe, 2007). I therefore adjusted this statistic by comparing with its expected value $\tilde{C}^3$. At each timestep t, I fit a linear regression between log-frequency and log-unique n-gram counts, and then computed the residual between the observed log count of unique trigrams and its expectation, $D^L = C^3_t - \tilde{C}^3_t$. The residual D^L, or *linguistic dissemination*, identified words with a higher or lower number of lexical contexts than expected.

Linguistic dissemination can separate words by grammatical category, as shown in Figure 5.1 where the mean D^L values were computed for words across common part-of-speech categories. Part-of-speech tags were computed over the entire corpus using a Twitter-based tagger (Gimpel et al., 2011), and each word type was assigned the most likely POS tag to provide an approximate distribution of tags over the vocabulary. Interjections had a lower median D^L than other word categories due to the tendency of interjections to occur in limited lexical contexts. Conversely, verbs had a higher median D^L due to the flexibility of verbs' arguments (e.g., subject and object of verbs may both be open-class nouns).

5.4 Results

The hypotheses about social and linguistic dissemination are tested under four analyses: correlation against frequency change in growth words; causal inference on probability of word growth; binary prediction of word growth; and survival analysis of decline words.

5.4.1 Correlational analysis

To test the relative importance of the linguistic and social context on word growth, I first correlated these metrics with frequency change ($\Delta_{f_t} = f_t - f_{t-k}$) across all growth words. This replicated the methodology in prior work by Altmann, Pierrehumbert, and Motter (2011) and Garley and Hockenmaier (2012), who analyzed word growth in several different internet forums. Focusing on long-term change with one year ($k = 12$) and two years ($k = 24$), I computed the proportion of variance in frequency change explained by the covariates using a relative importance regression (Kruskal, 1987).[6]

The results of the regression are shown in Table 5.4. All predictors had relative importance greater than zero, according to a bootstrap method to produce confidence intervals (Tonidandel, LeBreton, and Johnson, 2009). Frequency was the strongest predictor (f_{t-12}, f_{t-24}), because words with low initial frequency often showed the most frequency change. In both short- and long-term prediction, linguistic dissemination (D^L_{t-12}, D^L_{t-24}) had a higher relative importance than each of the social dissemination metrics. The social dissemination metrics had less explanatory power, in comparison with the other predictors and in comparison to the prior results of Garley and Hockenmaier (2012), who found 1.5% of variance explained by D^U and 1.9% for D^T at $k = 24$. The results were robust to the exclusion of the predictor D^L, meaning that a model with only the social dissemination metrics as predictors resulted in a similar proportion of variance explained. The weakness of social dissemination could be due to the fragmented nature of

[6]Relative importance regression implemented in the *relaimpo* package in R: `https://cran.r-project.org/package=relaimpo`

Table 5.4: Percent of variance explained in frequency change, computed over all growth words $\mathcal{G}$. $N = 26,880$ for $k = 12$, $N = 13,440$ for $k = 24$.

	Variance explained	Lower, upper 95%
f_{t-12}	10.8%	[10.2%, 11.5%]
D^L_{t-12}	0.584%	[0.461%, 0.777%]
D^U_{t-12}	0.307%	[0.251%, 0.398%]
D^S_{t-12}	0.120%	[0.085%, 0.191%]
D^T_{t-12}	0.246%	[0.171%, 0.379%]
f_{t-24}	21.4%	[20.4%, 22.4%]
D^L_{t-24}	1.29%	[1.05%, 1.64%]
D^U_{t-24}	0.400%	[0.346%, 0.493%]
D^S_{t-24}	0.287%	[0.201%, 0.392%]
D^T_{t-24}	0.272%	[0.226%, 0.380%]

Reddit, compared to more intra-connected forums. Since users and threads are spread across many different subreddits, and users may not visit multiple subreddits, a higher social dissemination for a particular word may not lead to immediate growth.

5.4.2 Causal analysis

While correlation can help explain the relationship between dissemination and frequency change, it only addressed the weak version of H2: it did not distinguish the causal impact of linguistic and social dissemination. To test the strong version of H2, I turn to a causal analysis, in which the *outcome* is whether a nonstandard word grows or declines, the *treatment* is a single dissemination metric such as linguistic dissemination, and the *covariates* are the remaining dissemination metrics. The goal of this analysis is to test the impact of each dissemination metric, while holding the others constant.

Causal inference typically uses a binary treatment/control distinction (Angrist, Imbens, and Rubin, 1996), but in this case the treatment is continuous. I therefore turned to an adapted model known as the *average dose response function* to measure the causal impact of dissemination (Imbens, 2000). To explain the procedure for estimating the average dose response, I adopt the following terminology: Z for treatment variable, X for

covariates, Y for outcome.[7]

1. A linear model is fit to estimate the treatment from the covariates,

$$Z_i \mid X_i \sim \mathcal{N}(\beta^\top X_i, \sigma^2). \tag{5.4}$$

   The output of this estimation procedure is a vector of weights $\hat{\beta}$ and a variance $\hat{\sigma}^2$.

2. The generalized propensity score (GPS) R is the likelihood of observing the treatment given the covariates, $P(Z_i \mid X_i)$. It is computed from the parameters estimated in the previous step:

$$\hat{R}_i = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{(Z_i - \hat{\beta}^\top X_i)^2}{2\hat{\sigma}^2}\right). \tag{5.5}$$

3. A logistic model is fit to predict the outcome Y_i using the treatment Z_i and the GPS $\hat{R}_i$:

$$\hat{Y}_i = \text{Logistic}(\hat{\alpha}_0 + \hat{\alpha}_1 Z_i + \hat{\alpha}_2 \hat{R}_i). \tag{5.6}$$

   This involves estimating the parameters $\{\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2.\}$ By incorporating the generalized propensity score $\hat{R}_i$ into this predictive model over the outcome, it is possible to isolate the causal effect of the treatment from the other covariates (Hirano and Imbens, 2004).

4. The range of treatments is divided into levels (quantiles). The *average dose response* for a given treatment level s_z is the mean estimated outcome for all instances at that treatment level,

$$\hat{\mu}(s_z) = \frac{1}{|s_z|} \sum_{z_i \in s_z} \hat{Y}_i. \tag{5.7}$$

   The average dose response function is then plotted for all treatment levels.

[7] Average dose response function implemented in the *causaldrf* package in R: `https://cran.r-project.org/package=causaldrf`

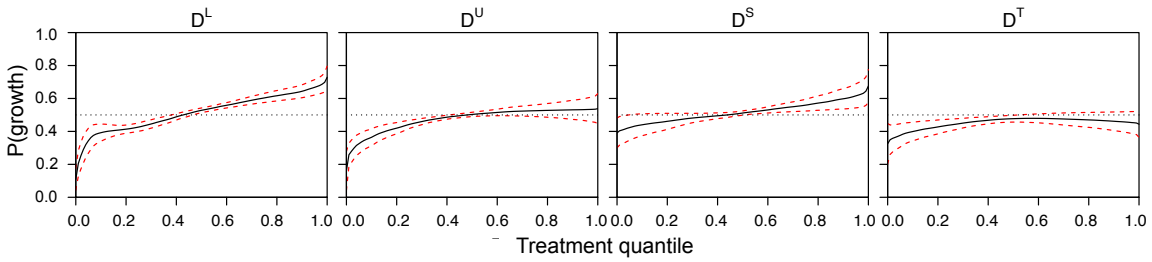


Figure 5.2: Average dose response function for all treatment variables, where outcome is probability of word growth. 95% confidence intervals plotted in red, chance rate of 50% marked with dotted black line.

Each dissemination metric was considered separately as a treatment. I considered all other dissemination metrics and frequency as covariates: e.g., for treatment variable D^L, the covariates are set to $[f, D^U, D^S, D^T]$. I bootstrapped the above process 100 times with different samples to produce confidence intervals. To balance the outcome classes, for each bootstrap iteration an equal number of growth and decline words was sampled.

The average dose response function curves in Figure 5.2 show that linguistic dissemination (D^L) produced the most dramatic increase in word growth probability. For linguistic dissemination, the lowest treatment quantile (0%-10%) yielded a growth probability below 40% (significantly less than chance), as compared to the highest treatment quantile (90-100%), which yielded a growth probability nearly at 70% (significantly greater than chance). This supports the strong form of H2, which states that linguistic dissemination is predictive of growth, even after controlling for the frequency and the other dissemination metrics. Subreddit dissemination also showed a mild causal effect on word growth, up to 60% in the highest treatment quantile. The other social dissemination metrics proved to have less effect on word growth.

5.4.3 Predictive analysis

I now turn to prediction to determine the utility of linguistic and social dissemination: using the first k months of data, can one predict whether a word will grow or decline in popularity? This is similar to previous work in predicting the success of lexical

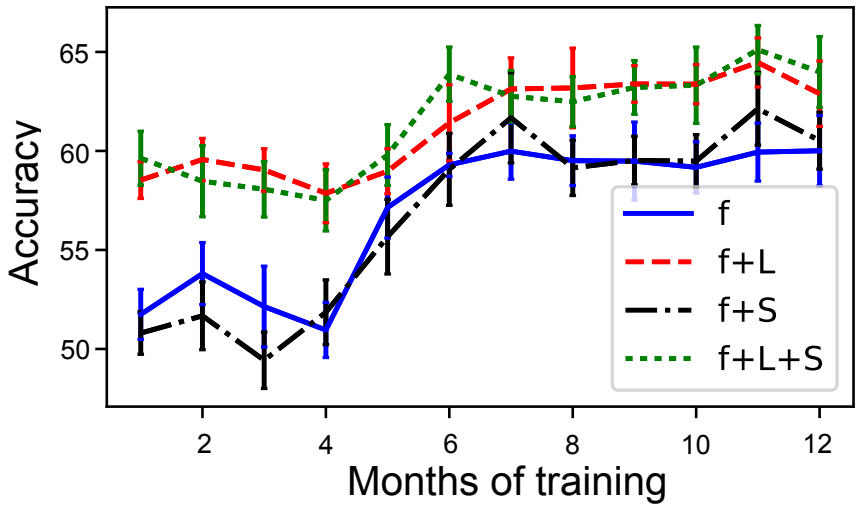


Figure 5.3: Prediction accuracy for different feature sets using $k = 1...12$ months of training data. f indicates frequency-only, $f + L$ frequency plus linguistic dissemination, $f + S$ frequency plus social dissemination, $f + L + S$ all features.

innovations (Kooti et al., 2012a), but the goal is to compare the relative predictive power of various dissemination metrics, rather than to maximize accuracy.

I used logistic regression with 10-fold cross-validation over four different feature sets: frequency-only (`f`), frequency plus linguistic dissemination (`f+L`), frequency plus social dissemination (`f+S`) and all features (`f+L+S`). Each fold was balanced for classes so that the baseline accuracy is 50%. Figure 5.3 shows that linguistic dissemination provided more predictive power than social dissemination: the accuracy was consistently higher for the models with linguistic dissemination than for the frequency-only and social dissemination models. The accuracies converge as the training data size increases, which suggests that frequency is a useful predictor if provided sufficient historical trajectory.

Part-of-speech robustness check Considering the uneven distribution of linguistic dissemination across part-of-speech groups (Figure 5.1), the prediction results may be explained by an imbalance of word categories between the growth and decline words. This issue is addressed through two robustness checks: within-group comparison and prediction.

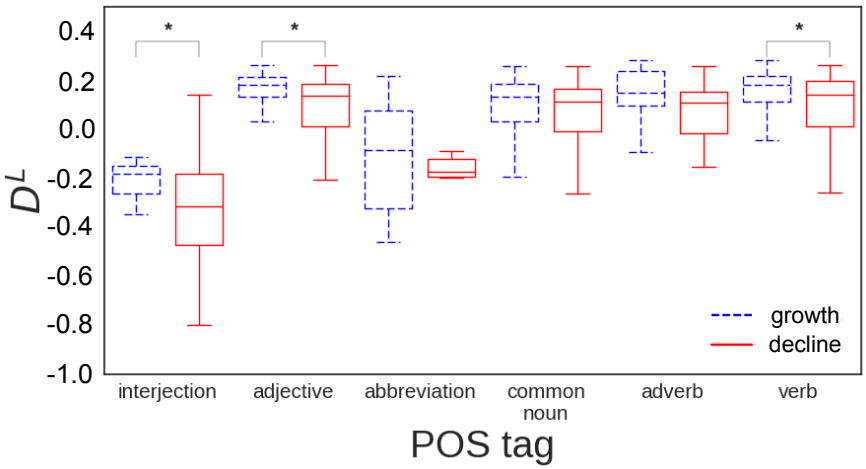


Figure 5.4: Distribution of D^L values across growth and decline words, grouped by part of speech tag. * indicates $p < 0.05$ in one-tailed t-test between growth and decline D^L values.

First, I compared the distribution of linguistic dissemination values between growth and decline words, grouped by the most common POS tags (computed in § 5.3.2). Each decline word was matched with a growth word based on similar mean frequency in the first $k = 12$ months, and their mean linguistic dissemination values during that time period were compared, grouped within POS tag groups. The differences in Figure 5.4 showed that across all POS tags, the growth words had a tendency toward higher linguistic dissemination with significant ($p < 0.05$) differences in the interjections, adjectives and verbs.

Next, I added POS tags as additional features to the frequency-only model in the binary prediction task. The accuracy of a predictive model with access to frequency and POS features at $k = 1$ reached 54.8%, which was substantially lower than the accuracy of the model with frequency and linguistic dissemination (cf. Figure 5.3).[8] Thus, linguistic dissemination thus contributed predictive power beyond what was contributed by part-of-speech alone.

[8]Higher k values yielded similar results.

5.4.4 Survival analysis

Having investigated what separates growth from decline, I now focus on the factors that precede a decline word's "death" phase (Drouin and Dury, 2009).

Predicting the time until a word's decline can be framed as survival analysis (Klein and Moeschberger, 2005), in which a word was said to "survive" until the beginning of its decline phase at split point $\hat{t}$. In the Cox proportional hazards model (Cox, 1972), the hazard of death λ at each time t is modeled as a linear function of a vector of predictors,

$$\lambda_i(t) = \lambda_0(t) \exp(\beta \cdot \mathbf{x}_i), \tag{5.8}$$

where $\mathbf{x}_i$ is the vector of predictors for word i, and β is the vector of coefficients. Each cell $x_{i,j}$ was set to the mean value of predictor j for word i over the training period $t = \{1...k\}$ where $k = 3$.

For words which began to decline in popularity in the dataset, I treated the point of decline as the "death" date. The remaining words were considered *censored* instances: they may have begun to decline in popularity at some point in the "future," but this time was outside the frame of observation. I used frequency, social dissemination and linguistic dissemination as predictors in a Cox regression model.[9]

The estimated coefficients from the regression are shown in Table 5.5. I found a negative coefficient for linguistic dissemination ($\beta = -0.330, p < 0.001$), which mirrored the results from § 5.4.2: higher D^L indicated a lower hazard of word death, and therefore a higher likelihood of survival. I also found that higher subreddit dissemination had a weak but insignificant correlation with a lower likelihood of word death ($\beta = -0.156, p > 0.05$). Both of these results lend additional support to the strong form of the hypothesis H2.

The predictive accuracy of survival analysis can be quantified by a *concordance*

[9]Cox regression implemented in the *lifelines* package in Python: `https://lifelines.readthedocs.io/en/latest/`.

Table 5.5: Cox regression results for predicting word death with all predictors (f+L+S) averaged over first $k = 3$ months. *** indicates $p < 0.001$, otherwise $p > 0.05$.

Predictor	β	S.E.
f	-0.207***	0.0492
D^L	-0.330***	0.0385
D^U	0.0053	0.0518
D^S	-0.156	0.0807
D^T	0.0825	0.0662

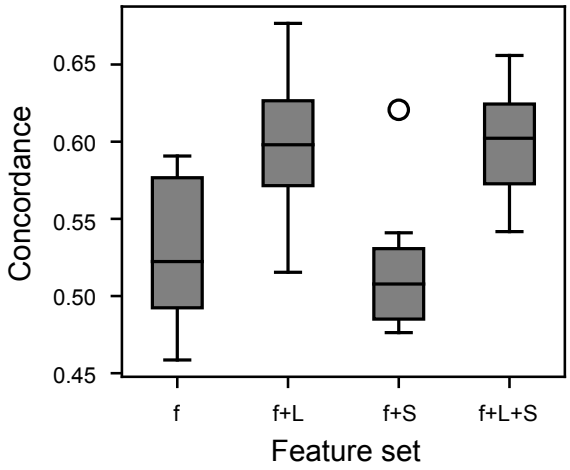


Figure 5.5: Distribution of concordance scores (10-fold cross-validation) of the Cox regression models across feature sets.

score. A score of 1.0 on heldout data indicates that the model perfectly predicts the order of death times; a score of 0.5 indicates that the predictions are no better than a chance ordering. I performed 10-fold cross-validation of the survival analysis model, and displayed the results in Figure 5.5.

The model with access to linguistic dissemination (f+L) consistently achieved higher concordance than the baseline frequency-only model (f), ($t = 4.29, p < 0.001$), and the model with all predictors f+L+S significantly outperformed the model with access only to frequency and social dissemination f+S ($t = 4.64, p < 0.001$). The result was reinforced by testing the goodness-of-fit for each model with model *deviance*, or difference from the null model. The f+L model had lower deviance, i.e. better fit, than the null model ($\chi^2 = 93.3, p < 0.01$), and the f+L+S did not have a significantly lower

deviance than the f+L model ($\chi^2 = 4.6, p = 0.80$), suggesting that adding social dissemination did not significantly improve model fit.

5.5 Limitations and future work

One limitation in the study was the exclusion of orthographic and morphological features such as affixation, which has been noted as a predictor of word growth (Kershaw, Rowe, and Stacey, 2016). Future work should incorporate these features as additional predictors. The study also omitted loanwords (cf. Chapter 7), unlike prior work in word adoption that has focused on loanwords (Chesley and Baayen, 2010; Garley and Hockenmaier, 2012). The early language-filtering steps eliminated most non-English words from the vocabulary, although it would have been interesting to examine loanword use in English-language posts. Finally, the study was limited by the focus on nonstandard words rather than memetic phrases (e.g., *like a boss*) which may show a similar correlation between dissemination, growth and decline (Bybee, 2006). On the social side, this study does not investigate different "levels" of social dissemination and assumes that all social units (users, threads, subreddits) should be treated equally regardless of network position. Tie strength could play an important factor: a word with high concentration in a small number of loosely connected subreddits may be more successful than a word that is weakly concentrated among many strongly connected subreddits (Milroy and Milroy, 1985).

From the methods perspective, I approximate linguistic dissemination using trigram counts, because they are easy to compute and they generalize across word categories. In future work, a more sophisticated approach might estimate linguistic dissemination with syntactic features such as appearance across different phrase heads (Kroch, 1989; Ito and Tagliamonte, 2003) or across nouns of different semantic classes (D'Arcy and Tagliamonte, 2015). Future research should also investigate more semantically-aware definitions of linguistic dissemination (Ryskina et al., 2020). The existence of semantic

"neighbors" occurring in similar contexts (e.g., the influence of standard intensifier *very* on nonstandard intensifier *af*) may prevent a new word from reaching widespread popularity (Grieve, 2018).

5.6 Contributions

- The study provides evidence that linguistic dissemination consistently explains word growth and decline, in a both causal and predictive tasks. In contrast, the metric of social dissemination that has proven useful in other work (Altmann, Pierrehumbert, and Motter, 2011; Garley and Hockenmaier, 2012) were less effective in this study as compared to the linguistic factor, suggesting a hierarchy of importance for factors in language change. This study reveals the importance of analyzing the linguistic properties of words rather than treating words as atomic units, in the context of word adoption. In terms of sociolinguistic theory, this study reveals the limits of assuming that speakers adopt words simply because they are socially well-evaluated, and it suggests that speakers rely more on the internal utility of a word when determining adoption or abandonment.

  The findings of this study can also inform more long-term studies in historical linguistics, where language changes may be started by external factors such as technological change and then moderated by internal factors such as syntax and semantics (Metcalf, 2004; Partington, 1993). For instance, the introduction of air transportation technology may lead to the adoption of a set of competing words related to e.g. airplanes, which speakers later differentiate based on their linguistic utility.

- The metric of linguistic dissemination can extend to different contexts or reformulated to capture different notions of context, e.g. studying grammatical context with POS tags. Due to its simple formulation, linguistic dissemination may

prove useful in similar situations involving large-scale user-generated text. This metric could be especially useful in predicting the outcome of situations in competing norms, such as emergent hashtags (Tsur and Rappoport, 2015), where norms that are more linguistically flexible stand more of a chance of success.

- In the context of social computing, this work speaks to the broader issue of understanding how innovations emerge over time. Analyzing how particular words become accepted by a community (Zhang et al., 2017) can reveal the community's values and the overall social structure of the community, e.g. if words need a small proportion of active members or a large portion of all members to be successful (different notions of social dissemination). This work can furthermore inform the study of other types of online innovations such as memes, where a text format that can apply to a wider variety of topical situations may be more successful than a format which simply appeals to many social groups (Rintel, 2013).

5.7 Thesis section summary

In this chapter, I test the relative influence of social and linguistic context in the adoption of new words, to address the open question of what structural factors drive language change. This chapter further demonstrates the utility of social media in testing open questions in sociolinguistics, due to the ability to compare multiple language changes happening in parallel which proves difficult in traditional analysis of spoken data or even typical written data e.g. newspapers (Chesley and Baayen, 2010). The broad range of topics available for discussion on social media provides a natural environment to observe words appearing a potentially wide range of linguistic contexts. Furthermore, the existence of multiple social levels in online media (community, thread, user) allows for a more thorough comparison of social evaluation that is less likely to be available in traditional spoken studies.

The next part of the thesis turns to another open sociolinguistics question which

concerns multilingual speakers, whose behavior in natural settings can be hard to scale. In the following two chapters, I leverage the longitudinal nature of social media to investigate the importance of social attitudes in explaining the language choices of multilingual speakers, specifically the choices to code-switch and the choice to use one form of a loanword over another.

CHAPTER 6

LANGUAGE CHOICE IN DISCUSSION OF A POLITICAL REFERENDUM

The third section of the thesis investigates the relevance of social attitudes with respect to multilingual language structure (RQ3): for multilingual speakers, how consistently do social attitudes explain their choice of which language to use in online discussions? Social attitudes are less easily defined than typical speaker-level attributes such as demographics but just as important in explaining a speaker's language use, because a particular language variety can be tied to larger political and societal issues (Auer, 2013; Blom, Gumperz, et al., 2000). In multilingual societies, the status of a minority language may reflect political marginalization of a sub-population who uses the language (Crameri, 2017; Moreno, Arriba, and Serrano, 1998), and the adoption of words from "outside" languages can be tied to larger cultural processes that privilege the status of outside languages (Low, Sarkar, and Winer, 2009; Thomason, 2001). To that end, social media provides a natural window into individual speakers' expressions of social attitude based on their sharing behavior (Gao et al., 2014), which may be more realistic than attitudes expressed through interviews or surveys. Therefore, the following section of the thesis takes a closer look at political and cultural attitudes in the context of code-switching and loanword use.

The following chapter investigates the choice of language in political discussions on Twitter, in the context of an independence referendum related to regional autonomy in Spain. I find that authors generally aligned with their attitudes using the expected language, e.g. that pro-independence authors used more of the minority language in order to signal their identity and personal connection to the issue. The study shows how minority language speakers can leverage the political status of their speech to stake a claim in political debates.

Note: Content for this chapter is drawn from Stewart, Pinter, and Eisenstein (2018).

This work was completed with the help of Yuval Pinter and Jacob Eisenstein.

6.1 Motivation

In a multilingual setting, an individual's preference to use a local language rather than the national one may reflect their political attitude, because the local language can have strong ties to cultural and political identity (Moreno, Arriba, and Serrano, 1998; Crameri, 2017). The role of linguistic identity is enhanced in extreme situations such as a referendum, where the voting decision may be driven by identification with a local culture or language (Schmid, 2001). In October 2017, the semi-autonomous region of Catalonia held a referendum on independence from Spain, where 92% of respondents voted for independence (Fotheringham, 2017). To determine the role of the local language (Catalan) in this setting, I apply the methodology used by Shoemark et al. 2017 in the context of the 2014 Scottish independence referendum to a dataset of tweets related to the Catalonian referendum.

I use the phenomenon of *code-switching* between Catalan and Spanish to pursue the following research questions in order to understand the choice of language in the context of the referendum:

- RQ1: Is a speaker's attitude toward independence strongly associated with the rate at which they use Catalan?
- RQ2: Does Catalan usage vary depending on whether the discussion topic is related to the referendum, and on the intended audience?

For the first question, the findings are similar to those in the Scottish case: pro-independence tweets were more likely to be written in Catalan than anti-independence tweets, and pro-independence Twitter authors were more likely to use Catalan than anti-independence Twitter authors (§ 6.3.1). With respect to the second question, I find that Twitter authors were more likely to use Catalan in referendum-related tweets, and that

they were more likely to use Catalan in tweets with a broader audience (§ 6.3.2).

6.2 Data

The initial set of tweets for this study, $\mathcal{T}$, was drawn from a 1% Twitter sample mined between January 1 and October 31, 2017, covering nearly a year of activity before the referendum, as well as its immediate aftermath.[1]

The first step in building this dataset was to manually develop a seed set of hashtags related to the referendum. Through browsing referendum content on Twitter, the following seed hashtags were selected: #CataluñaLibre, #IndependenciaCataluña, #CataluñaEsEspaña, #EspañaUnida, and CatalanReferendum. All tweets containing at least one of these hashtags were extracted from $\mathcal{T}$, and the top 1,000 hashtags appearing in the resulting dataset were manually inspected for relevance to the referendum. From these co-occurring hashtags, a set of 46 hashtags was chosen and divided into pro-independence, anti-independence, and neutral hashtags, based on translations of associated tweet content.[2] After including ASCII-equivalent variants of special characters, as well as lowercased variants, the final hashtag set comprised 111 unique strings.

Next, all tweets containing any referendum hashtag were extracted from $\mathcal{T}$, yielding 190,061 tweets. After removing retweets and tweets from authors whose tweets frequently contained URLs (i.e., likely bots), the final "Catalonian Independence Tweets" (CT) dataset comprised 11,670 tweets from 10,498 authors (cf. the Scottish referendum set IT with 59,664 tweets and 18,589 authors in Shoemark et al. 2017). 36 referendum-related hashtags appeared in the filtered dataset. They are shown with their frequencies (including variants) in Table 6.1 (cf. the 47 hashtags and similar frequency distribution in Table 1 of Shoemark et al. 2017).

To address the control condition, all authors of tweets in the CT dataset were

[1] A preliminary check of the data revealed that the earliest referendum discussions began in January, 2017.

[2] The authors of the original study had a reading knowledge of Spanish. For edge cases we consulted news articles relating to the hashtag.

Table 6.1: Hashtags related to the Catalonian referendum, their attitudes (neutral/pro/anti) and their frequencies in the CT dataset.

Attitude	Examples
Neutral	#1O (748), #1Oct (1351), #1Oct2017 (171), #1Oct2017votarem (28), #CatalanRef2017 (46), #CatalanReferendum (3244), #CatalanReferendum2017 (72), #JoVoto (54), #Ref1oct (90), #Referèndum (640), #Referendum1deoctubre (146), #ReferendumCAT (457), #ReferendumCatalan (298), #Votarem (954)
Pro-independence	#1ONoTincPor (18), #1octL6 (184), #CataloniaIsNotSpain (10), #CATvotaSí (3), #CataluñaLibre (27), #FreePiolin (293), #Freedom4Catalonia (2), #IndependenciaCataluña (9), #LetCatalansVote (3), #Marxem (102), #RepúblicaCatalana (212), #Spainispain (8), #SpanishDictatorship (9), #SpanishRepression (3), #TotsSomCatalunya (261)
Anti-independence	#CataluñaEsEspaña (69), #DontDUIt (12), #EspanaNoSeRompe (29), #EspañaUnida (4), #OrgullososDeSerEspañoles (55), #PorLaUnidadDeEspaña (2), #ProuPuigdemont (187)

collected to form a set $\mathcal{U}$, and all other tweets in $\mathcal{T}$ written by these authors were extracted into a control dataset (XT) of 45,222 tweets (cf. the 693,815 control tweets in Table 6 of Shoemark et al. 2017).

The CT dataset was very balanced with respect to the number of tweets per author: only four authors contribute over ten tweets (max = 14) and only 16 have more than five. The XT dataset also had only a few "power" authors, such that nine authors have over 1,000 tweets (max = 3,581) and a total of 173 have over 100 tweets. Since the results are macro-averaged over all authors, these few power authors should not significantly distort the findings.

Language Identification. This study compares variation between two distinct languages, Catalan and Spanish. I used the `langid` language classification package (Lui and Baldwin, 2012), based on character n-gram frequencies, to identify the language of all tweets in CT and XT. Tweets that were not classified as either Spanish or Catalan with at least 90% confidence were discarded. This threshold was chosen by manual inspection of

the *langid* output. In the referendum dataset CT (control set XT), `langid` confidently labeled 4,014 (56,892) tweets as Spanish and 2,366 (10,178) as Catalan. To address the possibility of code-mixing within tweets, one of the authors of the study and I manually annotated a sample of 100 tweets, of which half were confidently labeled as Spanish, and the other half as Catalan. We found only two examples of potential code-mixing, both of Catalan words in Spanish text.

6.3 Results

6.3.1 Catalan usage and political attitude

The first research question concerns political attitude: do pro-independence authors tweet in Catalan at a higher rate than anti-independence authors?

I analyze the relationship between language use and attitude on independence under two conditions, comparing the use of Catalan among pro-independence authors vs. anti-independence authors in (1) opinionated referendum-related tweets (tweets with Pro/Anti hashtags); and (2) all tweets. These conditions address the possibilities that the language distinction is relevant for pro/anti-independence Twitter authors in political discourse and outside of political discourse, respectively.

Method. The first step was to divide the Twitter authors in $\mathcal{U}$ into pro-independence (*PRO*) and anti-independence (*ANTI*) groups. First, the proportion of tweets from each author that include a pro-independence hashtag was computed as $\frac{N_{pro}^{(u)}}{N_{pro}^{(u)}+N_{anti}^{(u)}}$, where $N_{pro}^{(u)}$ ($N_{anti}^{(u)}$) is the count of tweets from author u that contain a pro- (anti-) independence hashtag. The *PRO* author set ($\mathcal{U}_{pro}$) included all authors whose pro-independence proportion was above or equal to 75%, and the *ANTI* author set ($\mathcal{U}_{anti}$) included all authors whose pro-independence proportion was below or equal to 25%. The counts of authors and tweets identified as either Spanish or Catalan are presented in Table 6.2.

To measure Catalan usage, let $n_{\text{CA}}^{(u)}$ and $n_{\text{ES}}^{(u)}$ denote the counts of Catalan and Spanish

Table 6.2: Tweet and author counts for the attitude study.

	Tweets with Pro/Anti hashtags		All tweets	
Group	*PRO*	*ANTI*	*PRO*	*ANTI*
# authors	713	242	1011	312
# Tweets	858	288	44,229	22,841

tweets author u posted, respectively. I quantified Catalan usage using the proportion $\hat{p}^{(u)} = \frac{n_{\text{CA}}^{(u)}}{n_{\text{CA}}^{(u)} + n_{\text{ES}}^{(u)}}$, computing the macro-average over each group $\mathcal{U}_G$'s members to produce $\hat{p}_G = \frac{1}{|\mathcal{U}_G|} \sum_{u \in \mathcal{U}_G} \hat{p}^{(u)}$. The test statistic is then the difference in Catalan usage between the pro- and anti-independence groups, $d = \hat{p}_{pro} - \hat{p}_{anti}$.

To determine significance, the authors were randomly shuffled between the two groups (pro/anti) to recompute d over 100,000 iterations. The p-value was the proportion of permutations in which the randomized test statistic was greater than or equal to the original test statistic from the unpermuted data.

Results. Catalan was used more often among the pro-independence authors compared to the anti-independence authors, across both the hashtag-only and all-tweet conditions. Table 6.3 shows that the proportion of tweets in Catalan for pro-independence authors ($\hat{p}_{pro}$) was significantly higher than the proportion for anti-independence authors ($\hat{p}_{anti}$). This accords with Shoemark et al. 2017, who found more Scots usage among pro-independence authors ($d = 0.00555$ for pro/anti tweets, $d = 0.00709$ for all tweets). The relative differences between the groups were large: in the all-tweet condition, $\hat{p}_{pro}$ is five times greater than $\hat{p}_{anti}$, whereas Shoemark et al. found a two-times difference ($\hat{p}_{pro} = 0.01443$ versus $\hat{p}_{anti} = 0.00734$ for all-tweet condition). All raw proportions were two orders of magnitude greater than those in the Scottish study, a result of the denser language variable used in this study (per-tweet code-switching vs. per-word code-mixing).

Table 6.3: Results of the attitude study. $d = \hat{p}_{pro} - \hat{p}_{anti}$.

	Tweets with Pro/Anti hashtags	All tweets
$\hat{p}_{pro}$	0.314	0.277
$\hat{p}_{anti}$	0.061	0.059
d	0.252	0.219
p-value	$< 10^{-5}$	$< 10^{-5}$

6.3.2 Catalan usage, topic, and audience

One way to explain the variability in Catalan usage is through *topic-induced variation*, which proposes that people adapt their language style in response to a shift in topic (Rickford and McNair-Knox, 1994). This leads to the second research question: is Catalan more likely to be used in discussions of the referendum than in other topics? This analysis was conducted under three conditions. The first two conditions compared Catalan usage in referendum-hashtag tweets (pro, anti, and neutral) against (1) all tweets; and (2) tweets that contain a non-referendum hashtag. This second condition was meant to control for the general role of hashtags in reaching a wider audience (Pavalanathan and Eisenstein, 2015a), and its results motivated the third analysis, comparing (3) @-reply tweets with hashtag tweets.

Method. I extracted all authors in $\mathcal{U}$ who have posted at least one referendum-related tweet and at least one tweet unrelated to the referendum into a new set, $\mathcal{U}_R$. Tweet and author counts for all conditions are provided in Table 6.4. The small numbers resulted from the condition requirement and the language constraint (tweets must be identified as Spanish or Catalan with 90% confidence). For an author u, I denoted the proportion of u's referendum-related tweets written in Catalan by $\hat{p}_C^{(u)}$, and the proportion of u's control tweets written in Catalan by $\hat{p}_X^{(u)}$. This analysis focuses on the difference between these two proportions $d^{(u)} = \hat{p}_C^{(u)} - \hat{p}_X^{(u)}$ and its average across all authors $\bar{d}_{\mathcal{U}_R} = \frac{1}{|\mathcal{U}_R|} \sum_{u \in \mathcal{U}_R} d^{(u)}$. Under the null hypothesis that Catalan usage was unrelated to

Table 6.4: Tweet and author counts for each condition in the topic/audience study. 'hash' stands for 'tweets with hashtags'.

Treatment set Control set	Ref. hash All tweets	Ref. hash All hash	Replies All hash
Authors	772	548	654
Treatment tweets	887	656	6225
Control tweets	31,151	13,954	10,319

Table 6.5: Results of the topic/audience study. $\bar{d}_{\mathcal{U}_R}$ is the difference in rate of Catalan use between treatment settings and control settings, averaged across authors.

Treatment set Control set	Ref. hash All tweets	Ref. hash All hash	Replies All hash
$\bar{d}_{\mathcal{U}_R}$	0.033	0.018	-0.031
Standard error	0.011	0.011	0.011
t-statistic	3.02	1.59	-2.79
p-value	0.002	0.111	0.005

topic, $\bar{d}_{\mathcal{U}_R}$ would be equal to 0, which I tested for significance using a one-sample t-test.

Results. The results, presented in the middle columns of Table 6.5, show that authors used Catalan at a significantly higher rate in referendum tweets than in all control tweets (first results column), but no significant difference was observed in the control condition where tweets include at least one hashtag (second results column). The lack of a significant difference between referendum-related hashtags and other hashtags suggests that the topic being discussed was not as important in choosing one's language, compared with the audience being targeted.

The second result reversed the prior finding that there were significantly *fewer* Scots words in referendum-related tweets than in control tweets (cf. Table 7 in Shoemark et al. 2017; $\bar{d}_u = -0.0015$ for all controls). This suggests that Catalan may have served a different function than Scots in terms of political identity expression. Rather than suppressing their use of Catalan in broadcast tweets, authors increased their Catalan use, perhaps to signal their Catalonian identity to a broader audience. This is supported by literature highlighting the integral role the Catalan language plays in the Catalonian

national narrative (Crameri, 2017), as well as the relatively high proportion of Catalan speakers in Catalonia: roughly 80% of the population has speaking knowledge of Catalan (Government of Catalonia, 2013), versus 30% population of Scotland with speaking knowledge of Scots (Scots Language Centre, 2011). There were also systemic differences between the political settings of the two cases: the Catalonian referendum had much larger support for separation among those who voted (92% in Catalonia vs. 45% in Scotland) (Fotheringham, 2017; Jeavens, 2014). These factors suggest a different public perception of national identity in the two regions within the context of the referenda, resulting in different motivations behind language choice.

Reply tweets

Earlier work has highlighted the role of hashtags and @-replies as affordances for selecting large and small audiences, and their interaction with the use of non-standard vocabulary (Pavalanathan and Eisenstein, 2015a). To test the role of audience size in Catalan use, I compare the proportion of Catalan in @-reply tweets against hashtag tweets.

Method. In this analysis, I took the treatment set to be all tweets made by authors in $\mathcal{U}_R$ which contain an @-reply but not a hashtag (narrow audience), and control against all tweets which contain a hashtag but not an @-reply (wide audience).

Results. The results in the rightmost column of Table 6.5 demonstrate a significant tendency toward less Catalan use in @-replies than in hashtag tweets. This trend supports the hypothesis that Catalan was intended for a wider audience. This effect may also be explained by a subset of reply tweets in political discourse being targeted at national figures, possibly seeking to direct the message at the target's followers rather than to engage in discussion with the target (in contrast to the minority language speakers studied by Nguyen, Trieschnigg, and Cornips, 2015). For example, one of the reply-tweets addressed a Spanish politician ("author1") in a conversation about a recent court case:

@author1 @author2 What justice are you talking about? What can a JUDGE like this impart?.[3] The same writer used Catalan in a more broadcast-oriented message: *Enough [being] dumb! We'll get to work and do not divert us from our way. First independence, then what is needed! Our part; #CatalonianRepublic*.[4] This provided a new perspective on the earlier finding by Pavalanathan and Eisenstein (2015a): by replying to tweets from well-known individuals, it may be possible to reach a large audience, similar to the use of popular hashtags.

6.4 Limitations and future work

One limitation of this study is the lack of geographic signals, because the sparsity of geotagged tweets prevented us from restricting the scope to data generated in Catalonia proper. Another potential limitation is that assumption that political hashtags are robust signals for political attitude. Other work has shown that political hashtags can be co-opted by opposing parties (Stewart et al., 2017b). On the methods side, I do not control for the content posted across languages which may relate to framing devices: e.g. it may be possible that Catalan-language posts tend to focus on voting, while Spanish-language posts tend to focus on the possible negative outcomes of Catalonian independence. I do not expect this to be the case, since both languages have the expressive capability to discuss the full range of political issues.

From the domain perspective, it may be difficult to directly compare Catalan and Scots due to different levels of mutual intelligibility with the respective native languages (Bailey, 1991; Wheeler, 1997), i.e. (written) Scots is likely farther from English than Catalan from Spanish. This concern is mitigated partly by the differences in the use of each language (word-level for Scots, sentence-level for Catalan) but should still be considered as a limitation that may reflect on future work, e.g. considering the use of even

[3] @author1 @author2 De que justícia hablas? De la que pueda impartir un JUEZ como este?

[4] Prou rucades! Anem per feina i no ens desviem del camí. El primer la independència, després el que calgui! El meu parti; #republicacatalana

less intelligible language pairs like Irish/English. Similarly, the relatively wider geographic spread of Catalan as a language as opposed to Scots (more rural/isolated; Scots Language Centre, 2011) makes it hard to compare directly the political value of both languages. It could be that Catalan speakers feel more comfortable using the language for "serious" discussions in general, including politics, while Scots speakers believe that a more serious audience is less likely to understand them; the choice between languages therefore may not reflect more profound social attitudes.

The findings of this study extend prior work on political use of Scots words on the inter-speaker level and Scots-English code-mixing on the intra-speaker level to examining language choice and code-switching, respectively. Further work is required to reconcile these results with prior work on topic differences and audience size (Pavalanathan and Eisenstein, 2015a). It may be the case that Catalan is useful across a variety of topic-related attitudes, such as national identity expressed through sports (Shobe, 2008). Future work may also compare the Catalonian situation with multilingual societies in which a minority language is discouraged (Karrebæk, 2013), or in which the languages are more equally distributed (Blommaert, 2011).

6.5 Contributions

- This study demonstrates the association of code-switching with political attitude, topic and audience, in the context of a political referendum. I corroborate prior work by showing that the use of a minority language was associated with pro-independence political sentiment, and I also provide a result in contrast to prior work, that the use of a minority language was associated with a broader intended audience. The first finding accords with work in sociolinguistic variation with attitude, and it suggests that speakers were aware of the political value of their language even to the point of modulating it in non-topical discussion. This furthers the understanding of social attitudes and language variation as relevant to not just

individual conversations but even to national discourse.

The second finding suggests that audience design depends not just on size but also on *type*, in this case whether the person is likely to respond to a comment. Such a distinction in audiences may be more relevant to political conversation, which is often more contentious (Demszky et al., 2019) and less dependent on typical formal/informal distinctions as compared to everyday conversation (Nguyen, Trieschnigg, and Cornips, 2015; Pavalanathan and Eisenstein, 2015a).

- For social computing methods, this study provides a content-agnostic view of political discussions that can be extended to other multilingual applications such as tracking responses to controversial discussions (Garimella et al., 2017). Rather than relying on the frequency of individual words, researchers can leverage the inherent differences in language choice among bilingual people to measure divergence of attitudes in online discussions. This can provide additional insight into political attitude in addition to the usual linguistic markers such as topical keywords (e.g. tracking pro-immigration phrases such as *open borders* alongside overall Spanish use among non-Spanish speakers).

CHAPTER 7
MORPHOLOGICAL INTEGRATION OF ENGLISH LOANWORDS IN SPANISH

In this chapter, I continue the study of social attitude in multilingual settings by investigating the integration of loanwords among Spanish speakers. Whereas the prior chapter focused on a minority language, I now investigate the adoption of words from an international *majority* language (English) and the relevance of speaker attitude toward the language's source culture.

The process of language mixing not only reveals differences in attitudes (Chapter 6), it also reveals long-term influences between languages. The adoption of English as a *lingua franca* around the world has helped spread individual words to many other unrelated languages (Chesley, 2010), to address concepts that originated among English-speakers such as *tweeting*. Studying how loanwords are accepted into other languages can reveal how speakers perceive the influence of other cultures on their native culture. Speakers who negatively perceive the loanword's source culture may also actively resist the acceptance of loanwords (Lev-Ari and Peperkamp, 2014), e.g. by using a native word equivalent or by explicitly marking the use of loanwords with different pronunciation. In contrast to language choice which carries social meaning in a variety of conversational contexts (Androutsopoulos, 2007; Gumperz, 1977), it is not well understood whether the integration of loanwords readily reflects a speaker's perception of the word's origin.

In this chapter, I conclude the study of social attitudes with a study of integration in English loanwords among Spanish-speaking authors on Twitter. I choose this context due to the abundance of English loanwords in Spanish (Gonzalez, 1999; Rodney and Jubilado, 2012) and the known influence of English-speaking culture, especially US American culture, on other countries (Crothers, 2017).

I find that loanword integration is more strongly associated with the "standard" language domain of newspapers, and therefore may be considered more formal. Furthermore, I find several important differences between how speakers tend to integrate loanwords versus native verbs, i.e. that integration is especially prominent among Spanish-speaking and Latin American authors. Lastly, I find that cultural attitude, measured by music consumption, has no significant effect on loanword integration. This suggests that the effect of culture on loanword perception may need to be measured with explicit metrics e.g. survey responses, and that loanword integration may not be as "marked" as language choice in terms of being tied to social attitudes (Myers-Scotton, 1998).

Note: this work was completed with the help of Diyi Yang and Jacob Eisenstein.

7.1 Motivation

Languages exchange loanwords constantly as multilingual people adopt words from one language to fill a gap in another (Poplack, Sankoff, and Miller, 1988). The English word *tweet* has been adopted by a number of other languages as a result of the success of the social media platform, e.g. producing the Spanish verb *tuitear*. Similarly to other forms of lexical change, the "success" of a loanword's introduction to a language depends on the word's similarity to the recipient language, the recipient language's lexical need, and cultural influence from the donor language (Calude, Miller, and Pagel, 2017; Garley and Hockenmaier, 2012; Zenner, Speelman, and Geeraerts, 2012). Modeling the adoption of loanwords can shed light on larger processes of cultural and linguistic change, such as the global reach of English into other major languages through borrowings (Pulcini, Furiassi, and Rodríguez González, 2012).

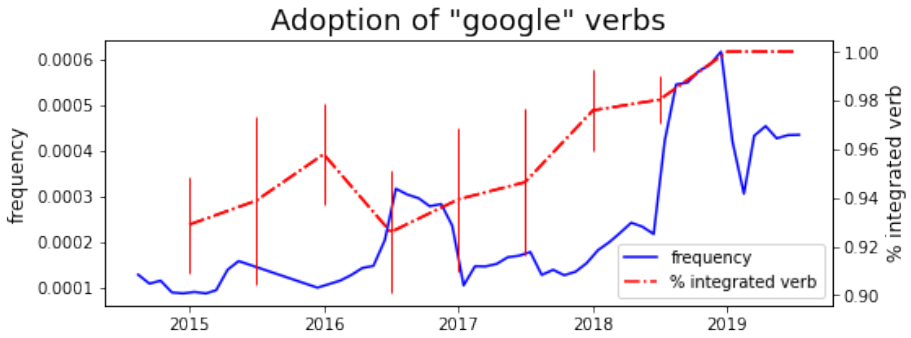
In addition to being borrowed, a loanword may also be *integrated* into a target language in terms of its pronunciation (Kang, 2011) and word structure (Poplack and Dion, 2012). For instance, the English verb *block* may be combined with Spanish

morphology to yield the integrated form *bloquear* ("to block") which contrasts with the typical paraphrase expression such as *hacer block* ("to do a block"). Loanwords may be integrated into the target language either gradually or instantly based on how well-attested the loanwords are at the time of use (Poplack, Sankoff, and Miller, 1988): a loanword that becomes more well-known in a community may have more pressure to integrate.

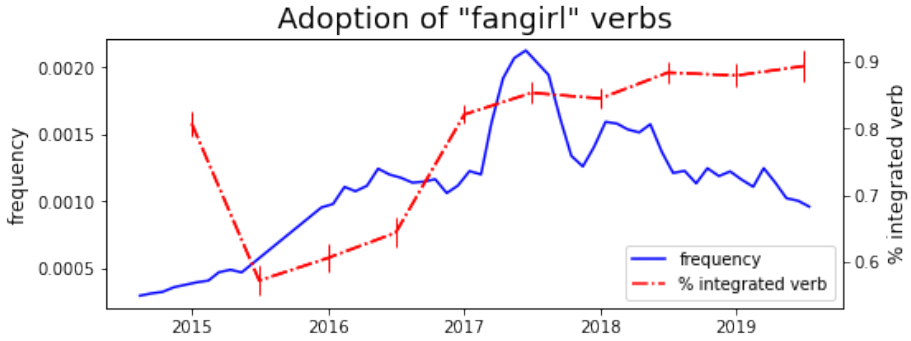
The issue of loanword integration is highly pertinent to Spanish, which has adopted a variety of loanwords from English due to language contact (Cacoullos and Aaron, 2003; Rodney and Jubilado, 2012). I show the trend of integration among a sample of English loanwords into Spanish in Figure 7.1, collected from a random sample of Twitter data (see § 7.2.3).[1] The well-known verb for *googlear* ("to Google") exhibits only a slight increase in its integration rate (92% to 100%), likely due to having been more well-known since the introduction of the search service in the early 2000s. In contrast, the newer verb for *fangirlear* ("to fangirl") exhibits a significant growth in its integration rate (60% to 90%), likely as a result of its sudden growth by an order of magnitude in tandem with the more recent adoption of the verb in English. Thus, morphological integration of loanwords may be dynamic in nature and not an inevitable process as previously posited (Poplack and Dion, 2012).

Loanword integration has been shown to be driven partly by language-internal factors such as frequency (Poplack and Sankoff, 1984) and lexical gaps (Zenner, Speelman, and Geeraerts, 2012). However, speaker-level factors also play a role in loanword production due to a speaker's cognitive and social constraints: speakers who are more familiar with the source language are less likely to feel the need to adapt the loanword (Haspelmath, 2009; Poplack, Sankoff, and Miller, 1988). Furthermore, a speaker who has a negative opinion of the loanword's source culture may also integrate the word more, to signal their affiliation to their native language (Hall-Lew, Coppock, and

[1]The rate of integration is equal to the number of verb tokens for a particular loanword that receive morphological integration, e.g. the number of tokens that are integrated as *fangirlear* normalized by the total number of verb phrases that represent "fangirl".



(a) "Google" verb pair: *googlear* and *buscar en Google* ("search on Google").



(b) "Fangirl" verb pair: *fangirlear* and *ser fangirl* ("be a fangirl").

Figure 7.1: Frequency and rate of verb integration over time, from 1% Twitter data sample from 2014-2019.

Starr, 2010; Lev-Ari, San Giacomo, and Peperkamp, 2014). Such individual-level factors may help explain the long-term integration of loanwords, particularly in socially concentrated systems such as online communities in which changes may occur rapidly (Danescu-Niculescu-Mizil et al., 2013).

One relevant speaker-level factor that relates broadly to sociolinguistic variation is media consumption: do speakers who consume media specific to the loanword's source culture tend to integrate the loanword differently? Consumption of media such as television and music has been shown to reflect patterns of language change such as dialect development (Androutsopoulos, 2014; Stuart-Smith et al., 2013). For bilingual people in particular, choosing one form of media over another (e.g. French music over English music) can index attitude toward a particular culture (Hernandez, 2010; Low, Sarkar, and Winer, 2009), as culture is constructed through concrete symbols such as art and traditions (Geertz, 2000).

I therefore propose to study loanword integration with respect to media consumption, with the goal of comparing its relative importance against known speaker-level factors in loanword integration. I investigate the integration of loanwords in social media and test a variety of social factors to determine potential sources of loanword integration among authors. Social media provides an ideal domain for such a study due to the tendency for wide-scale multilingual mixing that lends itself to borrowing between languages (Coats, 2018; Garley and Hockenmaier, 2012; Kim et al., 2014). First, I compare the rate of loanword integration in social media against a baseline corpus of Spanish newspapers to assess whether integration is associated with formality (Davies, 2020). Next, I investigate demographic factors and media sharing activity as social attributes that may explain verb integration, under the assumption that a speaker's cultural attitude can affect how they choose loanwords and therefore whether they choose to integrate a given loanword.

I consider the following research questions:

- RQ1: Does loanword integration correlate with writing domain (formal versus

informal)?

- RQ2a: On social media, what demographic and behavioral factors explain the use of integrated verbs among loanwords and native verbs?
- RQ2b: On social media, does media consumption explain the use of integrated verbs among loanwords and native verbs?

For RQ1, I find that authors of newspaper articles tend to use the integrated form of loanwords significantly more than social media authors, and that the newspaper articles are consistent in their level of integration regardless of location. The use of loanwords in newspapers may therefore have "levelled" to a strict writing standard, while on social media the integration of loanwords may be considered more flexible due to the lower standards for formality. For RQ2a, I find consistent evidence that Latin American authors and predominantly Spanish-speaking authors use more integrated loanword verbs and more integrated native verbs. For RQ2b, I find that media consumption plays a minimal role in loanword use and a surprisingly consistent role in native word use.

Taken together, these results demonstrate that loanword integration is likely a reflex of *formal style*: newspaper writers are more likely to adhere to a formal register (Biber and Conrad, 2019), and their consistent use of integrated forms suggests that integration for loanwords and native verbs is generally considered more formal. Spanish-monolingual authors consistently use integrated forms, which is to be expected of monolingual vs. bilingual speakers with respect to tendency toward more formal speech. Social media authors who share more Latin American media may also feel more aligned to Spanish language norms in general, which would lead them to use more formal style for native verbs due to the native verbs' relatively long entrenchment in the language (as compared to loanwords which are relatively newer to the language).

7.2 Data

7.2.1 Identifying loanwords

The use of a loanword is considered distinct from code-switching: whereas code-switching involves switching between languages, a loanword is a single word from a *donor* language that is produced within the same utterance as a *recipient* language (Poplack, Sankoff, and Miller, 1988; Cacoullos and Aaron, 2003). This study concerns the alternation between integrated verbs, i.e. those in which the loanword has been morphologically integrated into the language (*tuitear* "to tweet"); and light verbs, i.e. phrases in which the loanword is used as a noun (*poner un tweet* "to send a tweet"). The light verb phrases should be as semantically similar as possible to the integrated loanword verbs.

To identify a sample of valid words, a list of integrated loanword verbs was identified from two resources: Wiktionary and social media. First, I collected all verbs on Spanish-language Wiktionary that are identified as English-origin loanwords and end in the standard verb infinitive(*-ear*).[2] Using a sample of Reddit and Twitter data,[3] I then collected all words in Spanish-language posts[4] that match the structure `ENGLISH_WORD` + *-(e)ar*,[5] under the assumption that most loanword verbs are integrated using the *-(e)ar* conjugation (Rodney and Jubilado, 2012). The combined verbs were filtered to remove all cases of ambiguity: e.g. *plantear* can be formed by English *plant* + *-ear*, but it is a native Spanish word and therefore ambiguous.

For each loanword, I identified a corresponding light verb phrase that would have the equivalent semantic meaning as compared to the integrated form. Spanish has a closed

[2]Accessed 1 Jan 2020: `https://es.wiktionary.org/wiki/Categoria:ES:Palabras_de_origen_ingles`.

[3]Data sample ranges from 1 July 2017 to 30 June 2019. For Reddit this includes all comments, for Twitter this includes a 1% sample from the Twitter stream.

[4]Post language tagged using `langid`.

[5]English words collected from a standard spellcheck dictionary and filtered to exclude words shorter than $n = 4$ characters. Accessed 1 Nov 2019: `http://wordlist.aspell.net/dicts/`.

class of light verbs that speakers use to form phrases with nouns (Buckingham, 2013), e.g. *tomar un viaje* ("take a vacation") where *tomar* ("take") is the light verb. I used dictionary definitions from Wiktionary and WordReference to generate the majority of light verb forms, and I queried the internet for the remaining forms to determine their validity (e.g. comparing search results for *hacer un tweet* versus *poner un tweet*).

This process yielded 124 integrated and light verb pairs that I used to define the binary dependent variable of the study, i.e. integrated verb use versus light verb use. I show examples of the most frequent loanword and light verb pairs in Table 7.1.

Table 7.1: Top 5 most frequent loanwords and corresponding verb forms.

Loanword	Verbs	Count
Like	*likear*, *dar un like*	13,154
Connect	*conectear*, *hacer un conexión*	7857
Flip	*flipar*, *hacer flip*	6904
Stalk	*stalkear*, *ser un stalker*	5508
Tweet	*tweetear*, *poner un tweet*	5294

7.2.2 Identifying native verbs

Studying loanwords in isolation can yield interesting results, but it is important to determine whether the results represent loanword-specific trends or trends in verb integration in general. Similar to the other studies in the thesis, a control condition can help clarify the trends among loanwords. If loanword verbs are primarily used as integrated verbs by bilingual speakers, does this tell us about how the speakers treat loanwords or how they treat verbs in general?

To add the necessary control, I collected an additional set of integrated and light verb pairs that are native to Spanish. I first identified light verb constructions from several grammar blogs and dictionaries,[6] and I generated the corresponding integrated verb by adding a standard verb suffix (*-ar*) to the noun phrase and verifying with a dictionary. For

[6]E.g. "support verbs" mentioned here, accessed 1 Jan 2020: `https://comunicarbien.wordpress.com/2011/08/06/verbos-de-apoyo/`.

Table 7.2: Top 5 most frequent native word pairs and corresponding verb forms.

Native word	Verbs	Count
Dream	*soñar*, *tener un sueño*	39,392
Buy	*comprar*, *hacer la compra*	36,337
End	*terminar*, *poner término*	34,234
Use	*usar*, *hacer uso*	30,834
Test	*probar*, *poner a prueba*	29,930

the light verb construction *tomar un viaje* ("take a trip") with the noun *viaje*, I generated the integrated verb *viajar* ("travel").

This process yielded 49 native integrated and light verb pairs that serve as a baseline for the dependent variable of verb integration. For example, finding that a particular social variable explains integration among loanwords but *not* native words suggests that the variable is uniquely associated with loanword integration and therefore relates to multilingual ability. The most frequent native verbs and their translation can be seen in Table 7.2. Note the significantly higher counts in native verbs as compared to the loanword data, which more than addresses the imbalance in the number of word types i.e. fewer native verb types than loanwords.

I provide the complete list of loanword and native verbs in § B.1.

7.2.3 Collecting loanword data

For the main social media data of the study, I collected posts from a 1% Twitter sample between 1 July 2017 and 30 June 2019. All posts that contain at least one loanword verb form, either in the integrated form or light verb form, formed the main data for the study.[7] This yielded roughly 87,000 posts from 80,000 unique authors over the period of study.

Next, I collected all available prior posts from these loanword authors using both the original archive sample (2017-2019) and from the authors' full timelines (2014-2019).[8]

[7]I searched for the most frequent inflected forms of each verb, which include all forms of indicative present, simple future, simple past and imperfect. I also remove all verb forms that are ambiguous: e.g. the verb *acceso* ("I access") has the same spelling as the noun *acceso* ("access").

[8]Collected in Mar 2020, up to 1000 tweets from each author's timeline.

Some of the authors' timelines were unavailable due to e.g. account deletion, as a result of the difference in time between when the posts were written and when the extra data were collected. Despite this limitation, I recovered roughly 10 million posts from the authors, representing about 100 extra posts per author. I used this extra data to extract the necessary social variables for analysis.

7.2.4 Identifying cultural media

This study focuses on the relative influence of American/British media on Spanish speakers because media consumption may be reflect underlying cultural attitudes. Music is one form of media that people readily share online, that people can consume without understanding a given language, and that can mark cultural affiliation (Bryson, 1996; Hernandez, 2010; Way et al., 2019). A Spanish speaker may share music to show their connection to English-speaking culture, even if they do not speak English. I therefore use music as a proxy for cultural attitudes.

An author was considered to *consume* Spanish/Latin American media if they consistently shared links from artists known to from a Spanish/Latin American (*SLA*) background *more often* than artists known to be from a US/UK (*USUK*) background. To identify *SLA* artists and *USUK* American artists, I mined the super-categories established by Wikipedia that correspond to the different musician groups.[9] These lists were augmented with the "similar artists" suggestions from Spotify under the assumption that most Spotify users will tend to listen either to one group of musicians or the other, which will lead to consistent suggestions.[10] I removed all names that were likely to be ambiguous (e.g. Mario) and removed the intersection between the musician sets to avoid double-counting musicians (e.g. Luis Fonsi counts as both *SLA* and *USUK*). Let $\mathcal{U}$

[9] All sub-categories of these categories were queried on DBPedia (Jan 2020): `American_singers_by_genre`, `American_musical_groups_by_state`, `Singers_by_nationality` (for Latin American countries), `Latin_pop_singers`, `Latin_music_groups_by_genre`.

[10] The Spotify API provides suggestions for similar musicians (Accessed Jan 2020): `https://developer.spotify.com/console/get-artist-related-artists/`.

represent the *USUK* artists and $\mathcal{L}$ represent the *SLA* artists.

The links to songs on Spotify and YouTube provided a proxy for media consumption, under the assumption that both services are highly popular and are readily accessible to most internet users regardless of language. In a given post, I identified all URLs that are either a Spotify or a YouTube link. Next, for a given link I queried the corresponding API[11] to collect metadata about the song or video. For YouTube, the metadata included user-generated tags to help with searching indexing that often include artist names, and the video title often included an artist name in the format "TITLE - ARTIST."

I labelled a given YouTube video with a particular category (*SLA* vs. *USUK*) if: (1) it contained a user-generated tag that matches one of the artists in $\mathcal{U}$ or $\mathcal{L}$; or (2) the video title followed a typical "TITLE - ARTIST" format **and** the extracted artist matched one of the artists in $\mathcal{U}$ or $\mathcal{L}$. I labelled a given Spotify song with a particular genre if: (1) the artist name matches one of the artists in $\mathcal{U}$ or $\mathcal{L}$; or (2) one of the song's genres matches a typical *USUK* genre or *SLA* genre.[12]

To better understand the data, I show the most frequent artists in $\mathcal{U}$ and $\mathcal{L}$ extracted from the media data in Figure 7.2. The top artists were somewhat young, would mostly be considered "pop" music and were in wide circulation on radio stations during the time of the study (2017-2019).

7.2.5 Addressing confounds in media sharing

While media sharing can serve as a useful proxy for cultural affiliation, it is also subject to several limitations including the possibility for confounding. A person may share a link from Taylor Swift because they are a fan of her music, which could be confounded with their age as Taylor Swift's fans tend to skew younger (Katz, 2017). I therefore seek to

[11]Spotify API: `https://developer.spotify.com/documentation/web-api/`. YouTube API: `https://developers.google.com/youtube/v3/`. All APIs accessed between Feb 2020 and Mar 2020.

[12]I identified a genre as "typical" *USUK* if it tended to occur more frequently with artists in $\mathcal{U}$ than with artists in $\mathcal{L}$, and vice versa for *SLA* genres.

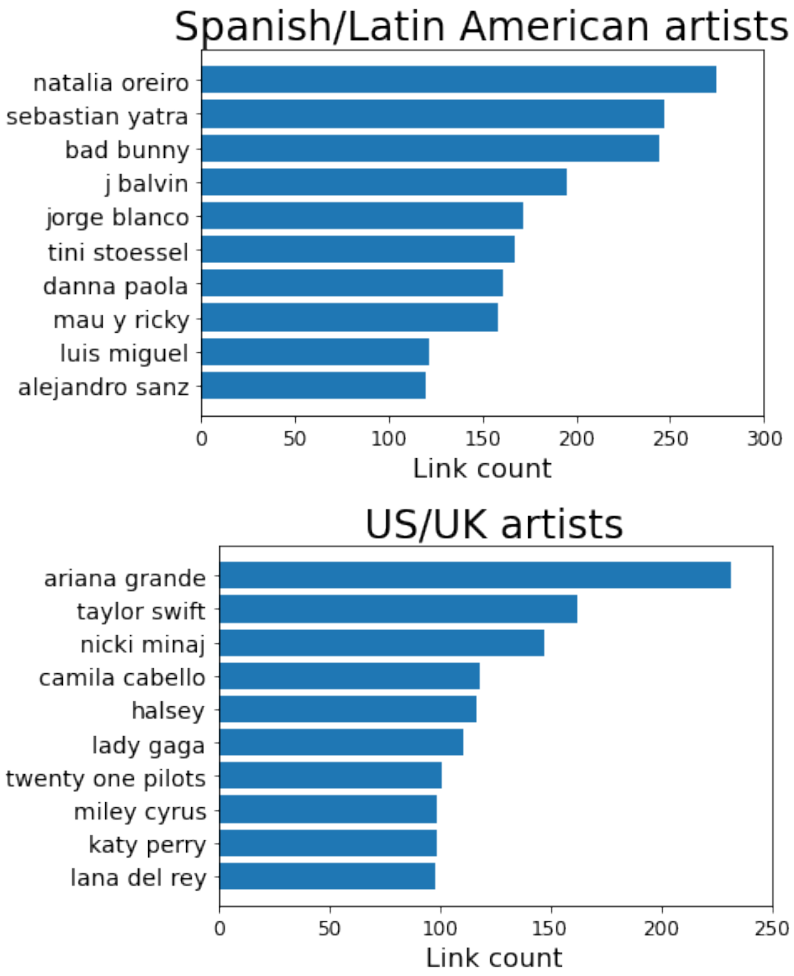


Figure 7.2: Top 10 artists in *SLA* and *USUK* categories.

balance the different media sources based on their likely audience ages.

The Facebook marketing API provides an estimate of audience sizes for sub-populations of users on Facebook, which is useful for companies to estimate the likely reach of a new advertising campaign. Researchers have leveraged the API to identify sub-populations that may otherwise be difficult or expensive to estimate, such as immigrants (Dubois et al., 2018; Stewart et al., 2019). Despite the self-selection bias in population (Facebook does not represent the full population), the API can provide a more organic, bottom-up estimate of interest-related sub-populations than many formal sources of data such as surveys.

In this study, I queried Facebook's marketing API to determine the *age* distribution of the artists' audiences. I collected the size of the audience of a given artist over different age categories: 13-25, 25-35, 35-45, 45-65+. These divisions likely captured generational divides in musical taste that would otherwise confound the analysis. I show examples of the "youngest" and "oldest" age distributions for musicians in the different categories in Figure 7.3. The age distributions matched the intuition about the fan bases, e.g. because Don McLean is older, his fans also tend to be older.

The age distribution of each media link was computed as the average over the age distributions of all artists who are contained in the link. Next, each media link was matched with a corresponding link from the opposite group with the lowest possible distance from the age distribution, i.e. for every *USUK* artist link, a corresponding *SLA* artist link with similar age distribution. While not optimal, this matching strategy improved the balance without seriously reducing the recall; the unbalanced and balanced age distributions shown in Figure 7.4 demonstrate a very close match in terms of media and variance.

In the rest of the analysis, I only counted a media link toward an author's sharing activity if the media link was matched through the procedure above.

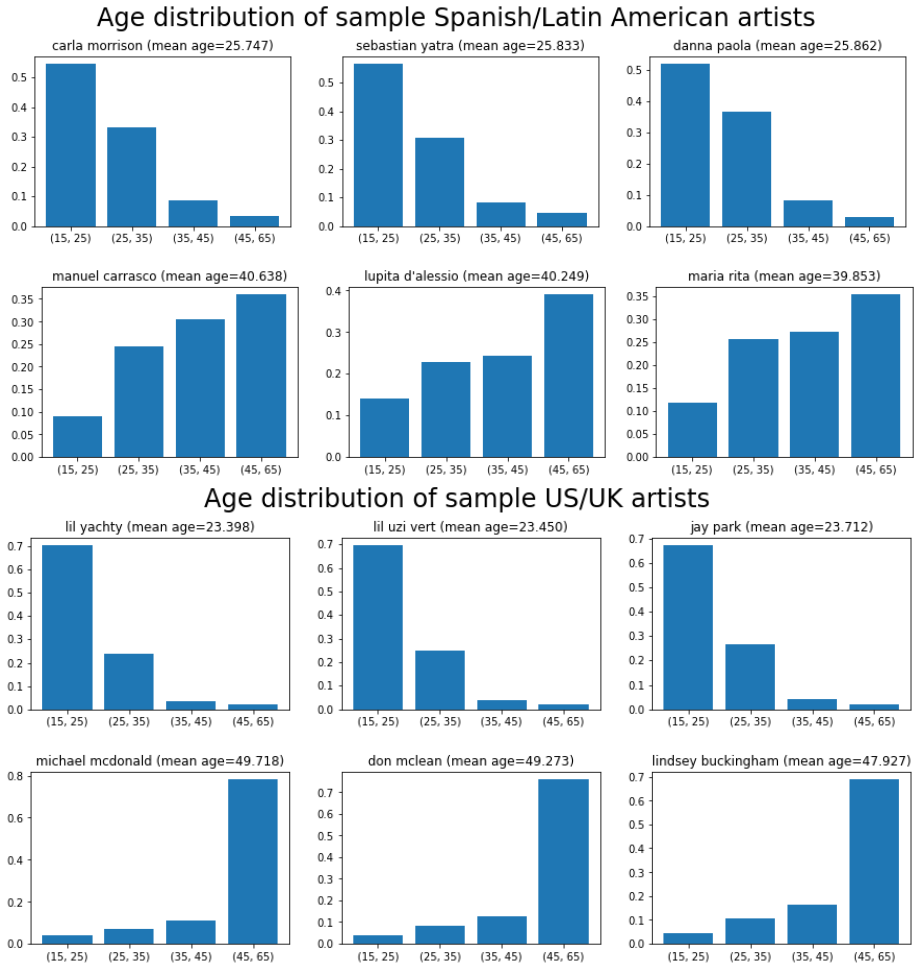


Figure 7.3: Audience age distributions for the "youngest" and "oldest" *SLA* and *USUK* artists, queried from Facebook marketing API.

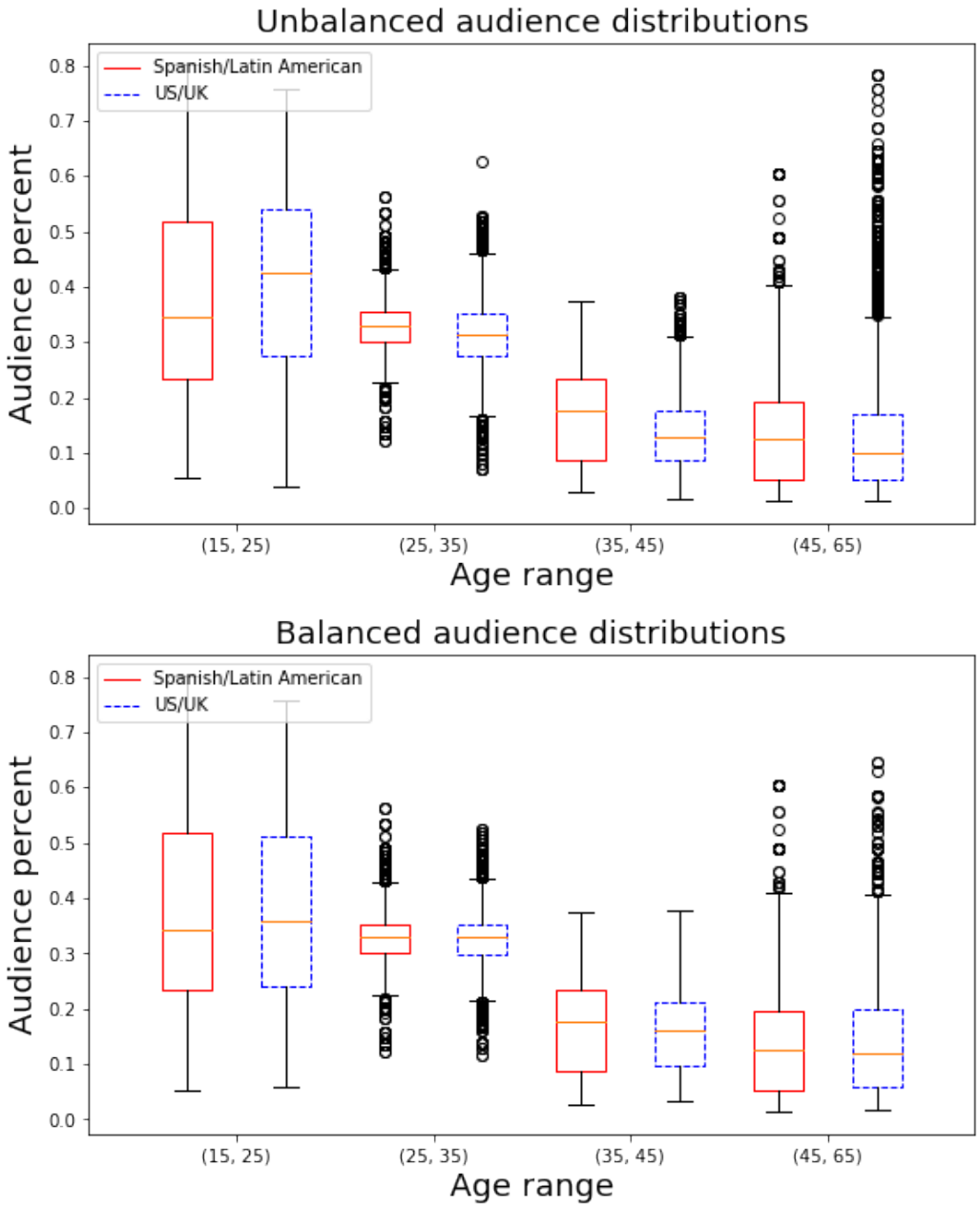


Figure 7.4: Unbalanced and balanced audience age distributions of *USUK* and *SLA* artists. The unbalanced age distribution exhibited a significant difference between the genres, while the balanced age distribution did not exhibit that difference.

7.2.6 Data variable summary

For RQ2, I seek to assess the relative importance of several individual-level factors in predicting loanword integration.

To address the question of demographic and behavioral factors in loanword use, I leveraged the following factors.

- **Activity**: authors who are more active online may accommodate more readily to internet-specific speech standards. I computed the author's mean number of posts per day, based on their prior activity.
- **Location**: the Spanish dialects spoken in Latin America have diverged significantly from Castilian Spanish (Lipski, 1994), which may result in different patterns of loanword adoption. Based on self-reported profile location, I identified authors' location[13] at the region level: Latin America, US, Europe, or other.
- **Language use**: bilingual speakers may be more likely to use the light verb forms of the loanwords, because they may rely on paraphrases to form verbs to address unfamiliar concepts (Jenkins, 2003), as compared to monolingual speakers who can access the standard integrated verbs more readily. I tagged all prior posts from all authors using `langid`,[14] and computed the rate of Spanish use for all authors who have written at least $N = 5$ posts, to identify consistent patterns of language use. I then binned language use under the assumption that language use may not be linear, using the bin [0-50%) for "low Spanish," [50-100%) for "medium Spanish," and [100%,) for "high Spanish." I used relatively high bin thresholds to accommodate the highly skewed distribution. I assumed that authors who use exclusively Spanish would be considered "strict" monolingual speakers as compared to more "relaxed" bilingual speakers (0-50%) or fully bilingual (50-100%) speakers.

[13] Following prior work in social media location detection (Kariryaa et al., 2018), I use an author's self-reported location in their profile as a location marker. I define an author as a resident of a particular country based on the presence of unambiguous country, state or city keywords in their profile location.

[14] I filter to posts with a confidence score above 90% to reduce likelihood of code-switching.

- **Verb use**: speakers who use more integrated native verbs, as opposed to the light verb forms, may be likely to use integrated loanword forms as well (assuming a common underlying mechanism for integration across verb types). I computed the rate of integrated verb use as the number of native integrated verb tokens (§ 7.2.2) normalized by the total number of native verb tokens produced by the author.
- **Sharing activity**: authors who share more content online may also be more connected to online norms in general and may therefore adopt the verb form that conforms most to the existing norms. I computed the rate of sharing as the percentage of posts that contain a URL ("link sharing") or retweet ("content re-sharing").
- **Media sharing**: authors who share *SLA* music more often may also be more closely aligned to Spanish-speaking culture more generally (Hernandez, 2010) and therefore choose the more formal verb form. I first computed the rate of music sharing as the proportion of links that contain a *SLA* artist, normalized by all links containing either a *SLA* or *USUK* artist. I then binned the media variable using [0,10%) for "low media," [10,50) "medium media," and [50,100) "high media." I used these bin sizes to accommodate the roughly bimodal distribution: authors tend to share only *USUK* media ($< 10\%$) or a high proportion of *SLA* media ($> 50\%$).

All variables in the study are summarized in Table 7.3. Note: while important demographic variables, I chose not to analyze each individual's gender and age due to the relative difficulty of extracting such information from social media data, particularly in non-English contexts (Wang et al., 2019b).

Table 7.3: Summary of all author-level variables used in study.

Variable type	Variable name	Description
General activity	Activity	Mean posts per day.
Demographics	Location	Author's geographic region based on self-reported location.
Language use	Language type	Percent of prior posts written in Spanish.
	Verb use	Percent of prior native verb posts that contain an integrated native verb (as compared to a light verb).
Sharing activity	Content re-sharing	Percent of prior posts that are retweets.
	Link sharing	Percent of prior posts that contain a URL.
	Media sharing	Percent of prior media-containing posts that contain *SLA* media (as compared to *USUK* media).

7.3 Results

7.3.1 Differences in integration by domain

The first hypothesis concerns the role of domain as a factor in loanword use. If newspapers are generally more formal than social media (Biber and Conrad, 2019), then we expect that loanwords and native verbs to be treated with the presumably more formal light verb forms.

To test this hypothesis, I collected additional data from a corpus of Spanish language newspapers from 21 different Spanish-speaking countries and regions.[15] While the specific guidelines for the newspapers studied are not readily available, I assumed that the newspapers' writers and editors tend to reinforce formal speech standards in general. I collected the top-50 most frequent loanword pairs and native verb pairs from the social media data, generated their conjugations as before and computed their raw frequencies from all different countries. For each pair of integrated verb and light verb I computed the rate of integrated verb use as the normalized frequency of the integrated verb. Formally, for a word base w, the set of all integrated verb forms $\mathcal{W}_{i,w}$, and the set of all light verb

[15]News On the Web Spanish, roughly 7 billion tokens total, accessed May 2020: `https://www.corpusdelespanol.org/now/`.

forms for the word $\mathcal{W}_{l,w}$, the rate of integrated verb use I_w is

$$I_w = \frac{\sum_{w_i \in \mathcal{W}_{i,w}} \text{count}(w_i)}{\sum_{w' \in W_{i,w} \cup W_{l,w}} \text{count}(w')}$$

The first key finding is that the rate of integration did not significantly differ for newspapers across locations. However, newspaper writers consistently used the integrated form of loanword and native verbs more frequently than the social media authors. Loanwords were integrated at a mean per-word rate of 89% in the newspapers as compared to 68% in social media, while native verbs had a rate of 92% in the newspapers and 80% in social media. I show in Figure 7.5 that social media writers consistently used integrated verbs at a significantly lower rate than newspapers across regions.[16]

The consistent difference between social media and newspaper writing, as well as the consistency across locations, suggests that the domain of newspaper writing has *standardized* the use of both loanwords and native words (Geeraerts, 2003). This seems to contradict prior corpus linguistic work that showed considerable creativity in light verb use in loanwords among Latin American newspaper writers (Buckingham, 2013). However, non-English newspapers are known to adopt English loanwords (Zenner, Speelman, and Geeraerts, 2012), especially those which matched a concept from English-speaking culture (e.g. *googlear* comes from the America-centric company Google). In the data, I found that the words with the highest differences in integration rates (newspaper - social media) tend to be related to technology, e.g. the integrated form of "link" (*linkear*) occurs at an 87% rate in newspapers as compared to 23% in social media (see rates for other example words in Table 7.4). This may be due to newspapers' imposing explicit norms on how to write about technology in a standard way (similar to the Royal Spanish Academy; Paffey, 2007), as compared to other lexical domains that may be less important to standardize.

[16] $p < 0.01$ comparing median rate of loanword integration and native verb integration across all location pairs, except Latin American ($p > 0.05$). I used the Wilcoxon test on rate of integration per-word and apply Bonferroni correction for multiple hypothesis testing.

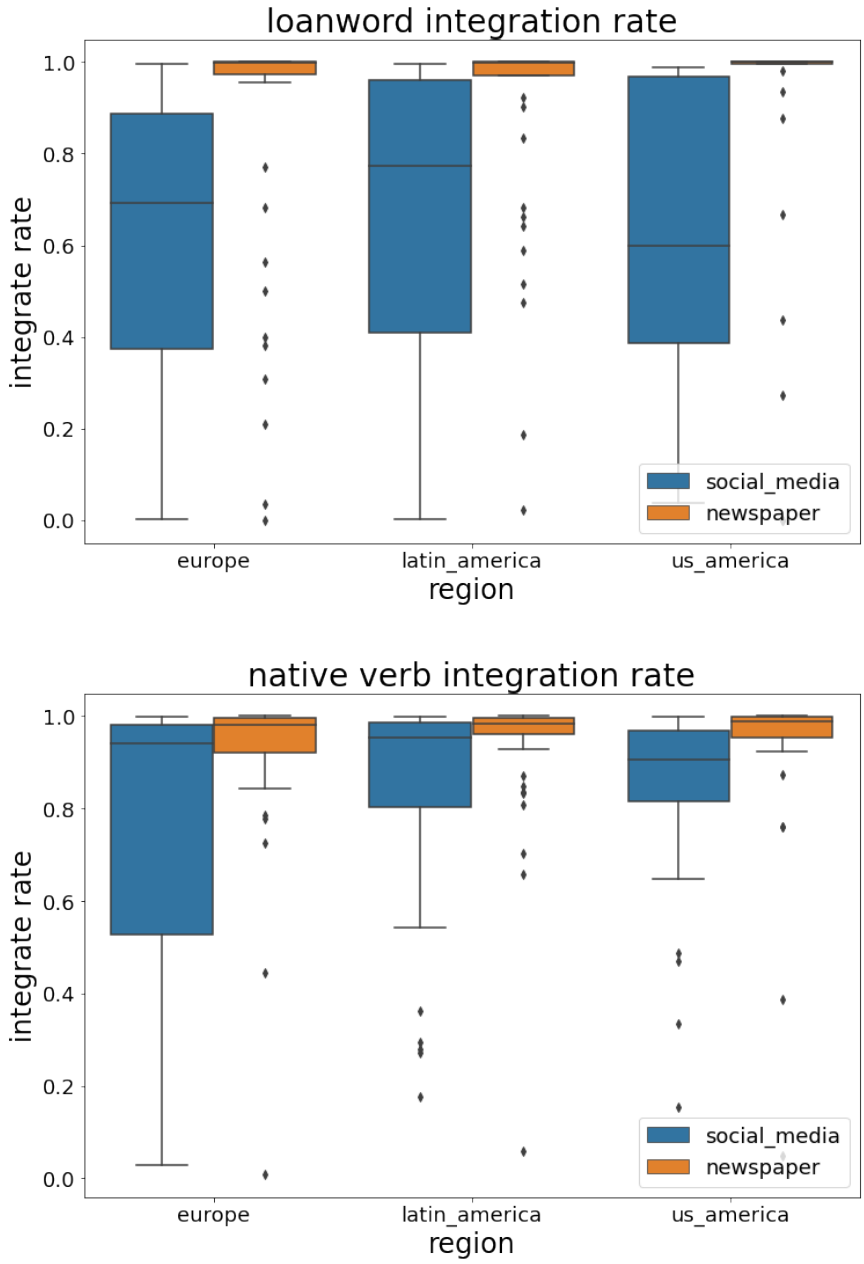


Figure 7.5: Integrated verb use across social media text (blue) and newspaper text (orange). Each point represents a single word.

Table 7.4: Loanwords with highest rate of integration difference between newspaper and social media writing.

Word	$I_{w,\text{social media}}$	$I_{w,\text{newspaper}}$	$\Delta\, I_w$
block	0.105	0.857	-0.752
hype	0.267	0.995	-0.728
link	0.227	0.872	-0.645
like	0.051	0.649	-0.598
perform	0.065	0.561	-0.496
access	0.523	1.000	-0.477
tweet	0.129	0.598	-0.470
boycott	0.593	0.989	-0.396
fangirl	0.632	1.000	-0.368
post	0.676	0.999	-0.323

In general, newspaper writers may more often defer to more morphologically succinct integrated verb constructions, similarly to how the progressive passive verb tense was systematically abandoned in English-speaking newspapers in favor of the more succinct (and "correct") active tense (Anderwald, 2014). Top-down writing guidelines from sources such as the Royal Spanish Academy (Amato et al., 2018), as well as prescriptive training among writers, can also contribute to a more uniform style regardless of the word's origin. Lastly, newspapers typically have multiple "layers" of writers who contribute to a given article (Bell, 1991b), and therefore tend to have layers of enforcement of writing standards, rather than social media where most accounts are controlled by a single author.

7.3.2 The role of demographics and behavior in integration

I now turn to individual-level prediction to assess the relative impact of different social factors (RQ2). If the use of integrated verbs is considered *more formal* as suggested by the prior analysis, then I should expect certain individual-level factors, such as higher Spanish use, to correlate with formality. I addressed this problem with logistic regression to predict the use of an integrated verb (1/0) for a given word token, using different subsets of author features specified in § 7.2.6 and a fixed effect for all sufficiently frequent authors

and word types.[17] To avoid possible overfitting among the fixed effect variables, I applied an L2 weight chosen to maximize likelihood on held-out data.[18]

To address RQ2a, I firsted test the role of demographics and behavior in the integration of loanwords and native verbs among social media authors. I show the regression results for demographics and author behavior in Table 7.5. I found the following significant results:

- **Location**: For both native words and loanwords, Latin American authors used integrated verbs at a higher rate. This relates to the divergence between Latin American Spanish dialects and other varieties: Latin American Spanish is known to have idiosyncratic light verb constructions (Buckingham, 2013) that may not combine with loanwords as readily as other dialects. Furthermore, Latin American authors may use integrated verbs more often due to their relative exposure to other languages. Since most Latin American countries use Spanish and they are surrounded by other Spanish-speaking countries, authors from Latin American may be more conservative in their language use as compared to authors from e.g. Spain, which is surrounded by populations that speak other languages. This difference may be a hypercorrection effect (DeCamp, 1972), such that the Latin American authors over-compensate for their perceived distance from Spain by using formal language more often. This would explain the opposite effect observed for European authors (less integrated verb use), as these authors may feel more confident in their Spanish use and more free to use less formal language.

- **Language**: For loanwords, high-Spanish authors used integrated verbs at a higher rate, and medium-Spanish authors used integrated verbs at a slightly higher rate. Integrated verbs could be considered canonical and therefore more accessible for monolingual speakers, while light verbs are more readily accessible to bilingual

[17] All authors and words with a count less than $N = 5$ were assigned to a `RARE` category to avoid sparsity.

[18] Weight selected from $\{10^{-5}, 10^{-4}, 0.001, 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.99\}$ to maximize held-out likelihood on a 10% test split of the data, for each separate regression.

Table 7.5: Regression results for loanword and native word integrated verb prediction. *** indicates $p < 0.001$, ** indicates $p < 0.01$, $\sim$ indicates $p > 0.05$.

		Loanword		Native word	
Variable	Variable bin	β	S.E.	β	S.E.
Intercept	-	-0.353	0.363	3.60***	0.135
Activity	-	-0.113***	0.0179	0.008	0.013
Location	Latin America	0.177***	0.026	0.124***	0.025
	Europe	-0.395***	0.051	-0.230***	0.0423
	US	-0.030	0.077	-0.041	0.070
	Other	-0.210	0.131	0.209	0.126
Language choice	High Spanish	1.990***	0.109	0.144	0.134
	Medium Spanish	0.837***	0.099	0.180	0.130
Integrated verb use	-	0.016	0.012	$\sim$	$\sim$
Content re-sharing	-	-0.612***	0.018	0.024**	0.013
Link sharing	-	-0.002	0.016	-0.019	0.013
Sample size		28,458	247,524		

speakers (González-Vilbazo and López, 2011). For example, the loanword phrase *dar un like* ("give a like") may sound more natural to a bilingual speaker who tends to use paraphrases when they are uncertain of the integrated forms' acceptability.

- **Sharing activity**: Authors who shared more content (via RTs) tended to use integrated verbs at a lower rate for loanwords and a higher rate for native words. This conflicts with the prediction that authors who share more will also be more likely to conform to online communication norms, due to their higher commitment to online activity. However, authors who share content more often may also be connected to a wider variety of other writing styles online through weak ties (Granovetter, 1973), which may in turn encourage more informal writing standards. One reason that this trend is reversed for native words could be the fact that native verbs are more entrenched in the language, and therefore they may have their formality enforced by people who have more consistent connections to online social life.

To summarize the findings for RQ2a, higher rates of integration among loanwords correlated with high Spanish use, Latin American location, and less content re-sharing.

The location results were shared between loanwords and native verbs, suggesting a similar underlying explanation for formality across the different domains.

7.3.3 The role of media consumption in integration

To address RQ2b, I added the media consumption variable described in § 7.2.4 and re-ran the same regressions as before. Including this variable reduced the size of the author population considerably, given that fewer authors in general tend to share music on social media. The results for the loanword and native word regressions using the media variable are shown in Table 7.6.

- **Location**: For native verbs, Latin American authors tended to use integrated verbs more often. This accords with the previous analysis. The lack of significant effects for the loanword verbs suggests that the sub-population of media sharers is not as affected by geography in their language use, possibly due to less emphasis on location as a form of relevant social cue.

- **Language**: Authors who use a medium amount of Spanish tended to use integrated loanword verbs more often. The lack of an effect for high-Spanish authors may stem from the data skew: among authors who share media, very few fall into the high-Spanish category.

- **Integrated verb use**: For loanwords, authors who tended to use more integrated native verbs also tended to use fewer integrated loanwords. This supports the previous finding (w.r.t. content sharing) that loanwords and native verbs have somewhat different constraints in terms of formality, and it may be the case that media-sharing authors treated the *light verb* form of loanwords as more formal. For example, if a media-sharing author states that they *like* a video (integrated verb), they may consider that usage less formal than *giving a like* (light verb) to the video, possibly due to age: use of YouTube and Spotify may be generally youth-affiliated,

Table 7.6: Regression results for loanword and native word light verb prediction, with media variable. *** indicates $p < 0.001$, ** indicates $p < 0.01$, $\sim$ indicates $p > 0.05$.

		Loanword		Native word	
Variable	Variable bin	β	S.E.	β	S.E.
Intercept	-	0.431	0.379	1.820	2.777
Activity	-	-0.252***	0.054	-0.008	0.032
Location	Latin America	0.021	0.117	0.164***	0.073
	Europe	-0.325	0.215	0.136	0.142
	US	-0.406	0.262	0.248	0.226
	Other	-0.185	0.429	-0.217	0.316
Language choice	High Spanish	0.735	0.660	0.011	0.427
	Medium Spanish	0.755***	0.276	0.636	0.262
Integrated verb use	-	-0.168***	0.059	$\sim$	$\sim$
Content re-sharing	-	-0.366***	0.051	0.012	0.034
Link sharing	-	0.004	0.053	0.025	0.033
Media sharing	High *SLA* media	0.108	0.109	0.126**	0.070
	Medium *SLA* media	-0.184	0.176	0.063	0.120
Sample size		1306		27,102	

regardless of the projected artist age distribution from Facebook.

- **Sharing activity**: For loanwords, authors who shared more content via RT generally used fewer integrated verbs. Assuming that light verb loanwords are more formal following the prior finding on integrated verb use, then sharing content more frequently may correlate with a stronger adherence to norms in general and therefore more formal verb use. The lack of an effect on the link sharing variable is explained by the fact that all authors in the analysis shared at least one media link, making them similar in their sharing habits.

- **Media sharing**: For native verbs, authors who shared more Spanish/Latin American music ("high *SLA* media") used integrated verbs more often. This confirms the hypothesis that more formal language use correlates with more Spanish culture alignment. However, there is no significant effect for loanwords, which suggests that cultural alignment (in its current form) only relates to formality among more well-established words i.e. native verbs.

- I note a consistent correlation between Spanish use and media sharing and therefore repeated the regression with only high-Spanish use authors. I find the same media effect held for native verbs, i.e. high *SLA* media authors used integrated verbs more often ($\beta = 0.139, p < 0.001$).

- It may be the case that *SLA* media sharing is related to a general cultural stance, such as linguistically conservative attitudes, which would explain the finding for native verbs. I tested this hypothesis by first separating the high-media authors from the low-media authors and then comparing the relative rates of URL sharing among the groups. Specifically, for each URL shared I computed the ratio of the proportion of shares among high-media authors and the proportion of shares among low-media authors: e.g. if `facebook.com` accounts for 50% of shares among high-media authors and 25% of shares among low-media authors, the ratio for `facebook.com` is $\frac{50}{25} = 2.0$. The top-50 and bottom-50 URLs by ratio were examined for the author groups, and I found that high-media authors tended to share more social media sites than low-media authors (11 unique sites among top-50 high-media URLs vs. 5 unique sites among top-50 low-media URLs; 2.89% of all high-media URL shares vs. 0.761% of all low-media URL shares; $Z = 34.6,\ p < 0.001$). This suggests that high *SLA* media authors were more open to sharing content from outside social media websites and therefore may have aligned more with general language norms (cf. the strong ties in social networks that support standard behavior; Granovetter, 1973), thereby supporting the use of integrated verbs.

To summarize the finding for RQ2b, media sharing did not explain the use of integrated loanwords but did explain the use of integrated native verbs, which suggests a difference in the level of awareness of formality in the different word groups. In general, the lack of effects in loanword integration may also be due to the significantly smaller sample population, leading to under-powered effects.

7.4 Limitations and future work

The main limitations of this study relate to data. The variable used to define media consumption leverages only music sharing habits, and I cannot rule out a possible effect of the media variable as defined in some other form, e.g. TV or film media content (Barnett, Oliveira, and Johnson, 1989; Sayers, 2014). In addition, the loanwords that provided the basis for the study tended toward more technological topics (e.g. *tweetear* is only used in the social media context), which may have in turn reduced the variety of speech communities represented in the work. I did not leverage all possible sources of data but I believe that the resources that I did use (e.g. dictionaries) provided a sufficiently wide representation of loanwords that was not overly specific to a particular speech community. Lastly, I make no claims of causality, and I leave to future work the more data-heavy question of whether listening to US American music over a long term generally causes a speaker to change their rate of loanword integration (cf. immigrant music changes in Way et al., 2019).

Future work in this space should consider other forms of linguistic integration that are readily available through written text, including spelling (*tweet* vs. *tuit*), adjective agreement (*ellos son cools* vs. *ellos son cool*), and noun pluralization (*los cowboys* vs. *los cowboy*) (Garley and Hockenmaier, 2012; Poplack, Sankoff, and Miller, 1988). Taken together, these varied forms of loanword integration may provide a more complete picture of a speaker's cultural attitude and willingness to obey or to reject their native language's rules. In another direction, the use of loanwords is a small slice of the larger picture of code-switching (Nguyen and Cornips, 2016; Solorio and Liu, 2008), and a speaker's use of integrated loanwords may be a single strategy in their broader repertoire of multilingual strategies. For instance, a multilingual speaker may tend to use loanwords in more Spanish-centric contexts such as in formal conversations, and then later use code-mixing among friends to signal closeness. In experimental settings, speakers may accommodate

their pronunciation of loanwords to one another (Lev-Ari and Peperkamp, 2014) which suggests that audience is a likely consideration for people in deciding to integrate loanwords morphologically. In terms of data, this study addressed only Spanish and should consider a wider variety of languages that have been influenced by English to test whether the findings on formality apply to all languages equally. Some countries such as Turkey may have explicit national policies to discourage loanword adoption in their native languages (Perry, 1985), which may in turn correlate with explicit aversion to morphological integration and therefore more light verb use. However, some languages such as Japanese may have reverse associations with loanword integration and formality (Tsujimura and Davis, 2011), i.e. treating integrated verbs such as *guguru* ("Google") as less formal than the light verb equivalent (*guguru suru*, "do Google").

7.5 Contributions

This study provides the following contributions to the thesis:

- I systematically compare the effect of register and individual-level factors that contribute to variation in the morphological integration of English loanwords in Spanish. The analysis among all authors reveals that loanword integration, like native verb integration, is primarily linked to formality as it is reflected in social constraints related to standard language usage, such as newspaper writing, high Spanish use, and Latin American geography. Addressing the attitude hypothesis, I find that social attitude as expressed through musical consumption does not explain loanword integration but correlates with native verb integration, i.e. more Spanish/Latin American media means more integrated verb use. This suggests that media consumption may be tied to formality in some contexts but not others, i.e. for native words that have had more long-standing acceptance among speakers.

  In terms of social attitudes, this chapter presents a sharp contrast with the political study. The negative result with respect to attitudes and loanword integration

suggests that cultural attitudes may not be as relevant to language choice as compared to political attitudes. A speaker's decision to use an integrated loanword may instead rely on attitudes that have more clear social values (Myers-Scotton, 1998), such as a speaker's ideas about English as a global language.

- The data collected in this study can be extended to other loanword studies for Spanish, such as comparing the relative influence of English on different Spanish dialects. The pairs of light/integrated verbs can serve as "seed pairs" to help researchers collect a broader range of paraphrases (Shoemark, Kirby, and Goldwater, 2018) for further study in the social construction of formality in Spanish. Future research can also extend the data collection pipeline to other languages with regular processes of integration. Specifically, this study's pipeline can be extended to any morphologically-rich language that uses integrated verbs and light verbs to accommodate English words, including Japanese (Tsujimura and Davis, 2011), French (Poplack and Dion, 2012) and German (Coats, 2018).

- For social computing, this work presents a method for quantifying and de-confounding media consumption to proxy cultural affinity expressed on social media. De-confounding is particularly applicable to studies of social media platforms where members tend to skew younger (Wojcik and Hughes, 2019), since researchers may want to assess consumption of cultural media across all age groups. This media metric can readily extend to other settings that involve cultural media consumption and may shed light on larger patterns of community norm development. In a sub-community online, people who tend to share "younger" media may also be those who drive the adoption of new norms in the community.

7.6 Thesis section summary

This chapter concludes the section of the thesis concerning the expression of social attitudes among multilingual people. I find that social attitudes provide more explanation for language variation when they are connected to clear social goals, such as making a political argument with a minority language, than when they are more likely part of general, unmarked conversation (i.e. music sharing). Furthermore, the loanword study provides evidence that what may appear to be a unique pattern may in fact be part of a broader tendency in language, i.e. a tendency for all integrated verbs to be associated with formality. Both studies demonstrate the importance of including a *control* condition (e.g. non-referendum discussion; native integrated verbs) when studying attitudes in language variation, to ensure that the observed pattern of variation is distinct from more general patterns in language.

CHAPTER 8
CONCLUSION

8.1 Thesis summary

This thesis addresses how regular patterns of language variation in online discussions can help answer open questions in sociolinguistic research. Specifically, I investigate how social media can address questions related to (RQ1) how people adjust to the communication expectations of their community and their discussion context, (RQ2) the relative influence of linguistic structure in word adoption, and (RQ3) the role of social attitudes in language choice among multilingual speakers. I address these questions through five quantitative studies of language variation on social media platforms.

8.1.1 RQ1 results summary

The following studies answer the first research question: how do speakers adjust their language to the assumed expectations of their community and their discussions, when they may not know the other participants?

1. In **Chapter 3**, I demonstrate the relevance of legitimate peripheral participation to the adoption of nonstandard word spelling in an online community related to pro-eating disorders. Community newcomers tend to drive the overall trend toward deeper variation in word spelling, even as they abandon such spellings during their own tenure in the community. Long-time members in the community tend to use deeper variants as well, suggesting that a person's intention to stay in the community is evident from their early posts. This study demonstrates how variation in word structure can reveal a tension between speakers' potential social goals (avoiding censorship and achieving legitimate participation in the community) that

can explain the overall change in the community.

2. In **Chapter 4**, I find that discussion participants actively adapt to their audience in discussion of crisis events, providing more or less information given their audience's likely awareness of the situation. The use of descriptor phrases reflects the collective attention paid to particular names during discussions and shows intuitive underlying social behavior, such as increasing descriptors in response to higher social engagement to prepare for a wider audience. Building on prior work (Staliūnaitė et al., 2018), I show that the relative time in an ongoing conversation (e.g. pre-event versus post-event) is just one factor among other social cues that can relate to a speaker's adaptation to their listeners.

With respect to RQ1, these studies collectively show that language variation can relate to a speaker's intention to adjust to their audience's expectations in order to share information and to participate in a larger community. Both of these forms of adaptation reflect a willingness among speakers to anticipate the needs of their listeners, even when the speaker is not guaranteed to know their listeners. Although sociolinguistics research has investigated similar situations such as radio broadcasts (Bell, 1984), social media provides insight into a variety of social situations. This variety can help reveal more nuanced behaviors, such as how speakers adjust to people who may be actively avoiding detection (Chapter 3) and to people who may have only recently joined the evolving discussion (Chapter 4). In fact, the second study reveals speaker behavior in contexts with more or less likelihood of audience awareness, showing that people adapt to a known-local audience (§ 4.3.1) and adapt to an newly emerging audience early in the crisis event (§ 4.3.2). These results across the spectrum of different conversation expectations reveal the value of social media into the spectrum of different audience configurations outside of the known/unknown binary (e.g. conversation among friends versus political speeches). Furthermore, similar configurations could be achieved experimentally (Rogers, Fay, and Maybery, 2013), but the naturally unfolding conversations in these studies provide a test

bed for language variation "in action," i.e. in the context of a person's everyday online experiences. The external "catalyst" of both censorship (Chapter 3) and rapidly-changing events (Chapter 4) provide a useful *in situ* example of how language variation can shift over time in response to actual, not simulated, social pressure.

In addition, both studies reveal the consistency with which people adapt to expected community *norms* in online discussions (Kraut and Resnick, 2012). In the case of the variant hashtags, the community norm toward evading censorship appeared to be obvious to newcomers, who used increasingly "deep" language variation as compared to normal hashtags that would be censored. In contrast, in the case of the descriptor phrases, the uncertainty around the unfolding crisis seemed to encourage speakers to maximize the likelihood of sharing useful information (i.e. rational communication; Grice, 1975), which was reflected by speakers who adapted to their static and dynamic audiences. The linguistic norms that online communities develop define their members and help to differentiate them from other communities (Kraut and Resnick, 2012): for instance, a community that values clear communication may adopt more strict language policies (Pavalanathan, Han, and Eisenstein, 2018) than other communities. Even seemingly lawless communities such as 4chan generally develop a shared understanding of how to participate, e.g. crude humor that borders on offensive (Bernstein et al., 2011). In the case of descriptor phrases, the addition of more contextual information may mark "other" status among listeners (e.g. non-locals) while less information can signal a shared social group identity among speakers and listeners that develops through the building of common ground (Acton and Potts, 2014; Doyle and Frank, 2015). Both Chapter 3 and Chapter 4 show that the expected norms of the discussions, explicitly defined by particular social constraints, are reflected by speakers' language choices. These studies address how language variation can help people collectively *construct* the expectations of their online spaces, which may be more malleable than the norms of offline communities. Far from arbitrary, the norms of the discussions encourage speakers to share information either with

a narrow, "hidden" audience or with a broad, wide-interest audience, both of which represent valid reasons for people to turn to social media in the first place.

The research in this thesis section helps reveal how people adapt the affordances provided by social media platforms to fit their own needs. In both studies, people adopt hashtags to address a particular type of discussion, and even within the same hashtag sub-groups may form to respond differently to the particular event. In Chapter 4, the non-active authors appeared to react more in sync with the collective ebb in attention by using fewer descriptors after the event, in contrast to the active authors who did not show such a trend. Similarly, in Chapter 3, there was no built-in mechanism for differentiating newcomers from old-timers, but the fact that hashtag use constituted a consistent set of practices that were non-trivial to learn seemed to provide incentive for natural differentiation among authors based on their adoption of the practices. In addition, the platforms' mechanisms for social feedback (e.g. likes, shares) provide concrete incentives for language choices. The active authors in Chapter 4 seemed to react to increased engagement by adding descriptor phrases to address a potentially more diverse audience, and the authors in Chapter 3 received slightly higher social engagement when using "deeper" variants. Authors' social reception in online settings may parallel the phenomenon of *backchanneling* (Mulac et al., 1998) where speakers reinforce each others' linguistic choices with cues like *uh-huh* during conversation, although backchanneling is often unconscious while online social reception is conscious. In online communication where speakers often do not meet face-to-face, such social feedback provides a signal to speakers that their listeners are engaged and find their language choices useful. The thesis demonstrates how these affordances for topical grouping and social feedback provide different incentives available to speakers in online discussions, when they make language choices.

8.1.2 RQ2 results summary

The following study answers the second research question: how readily do linguistic context dissemination and social context dissemination explain the adoption of words in online communities?

1. In **Chapter 5**, I find that linguistic dissemination consistently explains the adoption of nonstandard words in online communities, even more than the typical measure of social dissemination. In general, words with higher linguistic dissemination are more likely to grow and less likely to decline in the future, which suggests that speakers consider the potential utility of a word when considering whether to include it in their lexicon (Metcalf, 2004). This study shows the limitation of modeling innovation diffusion as a purely social process and reveals the utility of word context as a key variable in innovation dissemination.

This thesis section addresses a deep question in sociolinguistics, i.e. to what degree language change should be modeled with structural constraints (Labov, 1994) as opposed to social systems (Milroy and Milroy, 1985). While both factors are important, it is often difficult to compare them head to head because considering internal constraints requires either a specific subset of words in the same category (e.g. intensifiers; Partington, 1993) or an unusually large scale of spoken data that can highlight more than a few changes in progress (Tagliamonte and Denis, 2008). The high density of word changes available to study on social media permits the direct comparison of "successful" and "unsuccessful" words, which is a control condition often omitted from studies of word adoption (Kershaw, Rowe, and Stacey, 2016). By providing enough data for a kind of counterfactual condition, social media allows for paired treatment-control comparisons (§ 5.4.2) that reveal the full effect of individual factors on change in *general*, rather than just word growth. Social media furthermore provides a natural test-bed for the comparison of structural and social factors, as it permits a diversity of discussion situations in which

new words will compete for attention (Grieve, Nini, and Guo, 2016). By providing a top-down view of cross-thread and cross-community activity (Tan and Lee, 2015), Reddit provides a sufficiently diverse array of social factors to compare social structures comprehensively with internal factors. Similarly, the diversity of topics discussed on Reddit provide many opportunities for platform members to use words in variable or limited linguistic contexts, as compared to spoken data where new words could be so sparse as to only occur in the context of a single topic.

From the methods side, this study proposes a flexible metric for linguistic dissemination that has inherent validity, shown by the fact that part-of-speech categories can be differentiated by their linguistic dissemination. When internal structure is considered with respect to word adoption, it is often measured with more qualitative metrics, such as whether a word is perceived to fill a gap in the lexicon at a particular point in time (Grieve, 2018; Zenner, Speelman, and Geeraerts, 2012), or highly context-specific metrics, such as whether a word fits a particular grammatical role (Ito and Tagliamonte, 2003). Just as the social metrics are context agnostic, the linguistic metric should be generic to the specifics of a particular situation in order to guarantee that the results can be compared to other situations of change where language structure may be less important. Another useful aspect of this study's linguistic metric is its light-weight nature, as one only needs to compute trigram counts and frequency instead of more dense computation such as word embeddings (Kulkarni et al., 2015). In order to address large-scale language change, researchers should consider language-internal metrics that are agnostic to the specific type of change and can scale without excessive computing needs.

8.1.3 RQ3 results summary

The following studies answer the third research question: for multilingual speakers, how consistently do social attitudes explain their choice of which language to use in online discussions?

1. In **Chapter 6**, I find consistent evidence for the expression of political attitude in the choice of languages on social media, during discussion of an independence referendum in Spain. Bilingual people tend to use their minority language more often when they are pro-independence and when they are discussing the referendum itself, which suggests an active construction of political attitude through language choice that is more pronounced than in prior work (Shoemark et al., 2017). This study demonstrates how language choice can help individuals share their political identity, particularly in "extreme" situations related to concrete political consequences (e.g. independence). For the thesis, the study shows how social media can reveal the nuances of language choices by tracking a speaker's behavior in different contexts with respect to the same issue.

2. In **Chapter 7**, I first frame the question of loanword integration (e.g. English *tweet* to Spanish *tweetear*) as the alternation between integrated verbs and light verbs (*to tweet* versus *to send a tweet*). I show that the process of loanword integration is likely related to *formality*, based on the fact that newspapers use integrated verbs at a higher rate than social media authors. Furthermore, loanword integration is explained less by cultural factors such as media consumption and more by demographic factors, such as high Spanish use and Latin American location. Therefore, the use of integrated loanwords among multilingual people may have less explicit cultural meaning than anticipated based on prior work in phonetic integration (Lev-Ari and Peperkamp, 2014). Furthermore, this study tests the alternation between integrated verb and light verb, which provides a new perspective on loanword adoption that complements the typical approach that tests a loanword's similarity to the donor/recipient languages (e.g. testing whether loanword is pronounced closer to the donor or recipient language; English *tweet* versus Spanish *tuit*).

With respect to RQ3, these studies demonstrate that social attitude is not always

reflected in multilingual speakers' language use, and that this connection may be moderated by the social *meaning* of a given language variable (Eckert, 2016). To be clear, this thesis addresses a subset of social attitudes that were assumed to correspond to non-linguistic behaviors (politics and music), as opposed to other attitudes such as language ideologies that directly reflect the value of a particular variety (e.g. negative stigma toward rare dialects; Preston, 2002). However, the social attitudes studied in this thesis reflect the broader idea of sociolinguistic *identity* (Bucholtz and Hall, 2005) and therefore how people present themselves to others (Goffman, 1978). Prior sociolinguistic work holds that a speaker's identity is constructed during conversations rather than existing as a static psychological construct (Eckert, 2008), and this thesis investigates this construction process in multilingual speakers whose identity is marked as distinct from monolingual speakers every time they switch languages.

Focusing on attitudes as a component of identity, the studies in this section provide evidence for the relative difference between attitudes linked with clear social value, e.g. political affiliation associated with a particular voting behavior (Hall-Lew, Coppock, and Starr, 2010), and more ambiguous attitudes, e.g. affinity with a particular culture (Low, Sarkar, and Winer, 2009). An attitude that relates to definite action and clear position within society (i.e. politics) may be reflected in language variation more readily than a more ambiguous attitude. In contrast to political views, a person may share music on social media simply because they liked the musical aspect, rather than the underlying cultural values. This is particularly relevant to self-presentation online, where the "context collapse" of different conversations may mean that a person's attitude may be understood by some and misinterpreted by others (Marwick and boyd, 2011). In the context of studying language variation, it may be best for researchers to consider attitude expression from the listener side rather than the speaker side, to consider what impression the speaker is making versus what their possible intentions were (cf. impressions *given* versus impressions *given off*; Goffman, 1978).

In addition to the inherent social value of speaker attitudes, the thesis provides insight into how attitudes interact with other aspects of a speaker's background. In Chapter 7, the fact that cultural attitude explains loanword integration less well than speaker demographics may indicate that verb integration is tied more strongly to underlying societal norms, e.g. regional differences between Latin America and European speakers. In Chapter 6, I find that Catalan speakers tend to use the minority language more often in broadcast messages, and these speakers appear sufficiently confident in the social value of the minority language (Government of Catalonia, 2013) that they share it even with an audience who may not be able to translate it (due to the English or Spanish hashtags used). These interactions between different aspects of identity are particularly important when considering multilingual people, who may perceive themselves as managing multiple identities in conversation based on their social connection to different languages (Auer, 2013; Christiansen, 2015). The work in this thesis shows that social attitude should be considered as one part of a larger repertoire of speaker identity among multilingual people, which have varying influences on their language variation based on particular situations. For example, political attitude may be more relevant during discussions of minority-majority language dynamics (Blommaert, 2011), while discussions relating to international differences may bring out cultural attitudes. Blom, Gumperz, et al. (2000) showed that the introduction of linguistic outsiders into a community can lead to the original population switching to their native language, which supports the idea that certain social situations bring out the relative value of language choices more than others.

In terms of domain, social media provides a window into the wide variety of resources that people use to construct their identity. This is particularly valuable to understanding media sharing habits (Johnson and Ranzini, 2018), as social media provides a "wider net" to catch media sharing as compared with the usual survey approach (Thomas, 2017) that requires a researcher to know the relevant media beforehand. This approach proved useful in both Chapter 6 and Chapter 7, as it covered a

wide range of political and musical expression that many outsider researchers may not have considered. The constructs used to identify attitudes are not restricted to multilingual situations on social media but can apply to similar scenarios to investigate the social motivation of language variation, such as the use of political slang words (Hossain, Tran, and Kautz, 2020) to express strong liberal or conservative attitudes. Furthermore, while the specific definition of cultural attitude in Chapter 7 (media sharing) was not as relevant to the specific situation studied, the construct may certainly play a role in other contexts where the language variation is more "marked" (Myers-Scotton, 1998), including the distinctive syntax of African American English (Wolfram and Thomas, 2008) which is often used in American music (Cutler, 1999; Eberhardt and Freeman, 2015). Therefore, the construction of social attitude through sharing behavior, as opposed to self-reported beliefs, may prove useful even beyond the current scope of multilingual speakers.

8.1.4 Overall summary

This thesis demonstrates the value of social media and online discussions in addressing open sociolinguistic questions that are otherwise difficult to address at scale. Specifically, all of the studies considered insight into the *why* of language variation, which is under-explored among computational sociolinguistics studies that typically focus on *what* types of variation manifest among different populations (e.g. geographic differences).

In terms of findings, adaptation to conversation expectations (RQ1) and expression of attitudes (RQ3) represent broader patterns of *participation* in online discussions (Kraut and Resnick, 2012), by which people actively contribute and benefit from discussion instead of e.g. lurking (Preece, Nonnecke, and Andrews, 2004). Studying speakers' adaptation to conversation expectations and attitude expression on social media reveals how readily people adapt their behavior to online spaces, albeit with slightly different affordances. Multilingual people in particular exhibit a considerable amount of cross-cultural and cross-political self-expression on social media (RQ3), which may be

helped by the semi-anonymous nature of online discussions (Peddinti, Ross, and Cappos, 2014) that allows for greater freedom of expression. In addition, multiple studies in this thesis have revealed the complicated nature of how audiences in online environments may be expected to respond to speakers' behavior. In a space where potentially many people may view a person's posts (Marwick and boyd, 2011), speakers need to consider the tradeoff between customizing their message (narrow audience) and increasing the likelihood of a response (broad audience). This applies especially to Chapter 4 and Chapter 6, where a speaker's choice of audience may be more high-stakes due to pressure from unfolding external events. As people continue to split their social lives between offline and online spaces (Ploderer, Howard, and Thomas, 2008), researchers will need to determine the differences in how people adjust to their listeners and express their identity with more or less explicit social mechanisms at work (e.g. uncertain audience composition). Finally, the thesis provides differing views of the social *evaluation* of speakers' participation in online discussions. The studies in RQ1 found consistent correlation with social engagement and language variation that fit the expectations of the other conversation participants, but the study in RQ2 found that social evaluation played a limited role in explaining the growth and decline of words over time. While social evaluation may reasonably relate to specific cases of language variation, it may be less pertinent to more general cases of change, since language change may have more or less social value attached depending on the particular case. For instance, the laughter words *haha* and *hehe* may have different connotations in a particular conversation, but across Reddit in general they may have roughly the same social meaning, which may mean that social dissemination also has less predictive power for the growth or decline of these words.

More to the point of the original thesis statement, the studies presented here have shown that social media can shed light on long-standing questions in sociolinguistics by providing a more comprehensive view into social factors that are difficult to assess in

offline settings. The most critical benefit of social media is the wide variety of linguistic patterns that would be rare or contrived in most corpora of spoken conversations. Discussions on social media are certainly moderated (Chandrasekharan et al., 2017), but the relatively free-form conversations on platforms such as Twitter and Reddit permit a wide variety of topics and language use. This is important for nonstandard words (Chapter 5) and loanwords (Chapter 7) which are notoriously rare in spoken conversation (Poplack, Sankoff, and Miller, 1988). Having access to a large sample of words provides not just a high token count but also a high *type* count (e.g. over 1000 growth words in Chapter 5), which reduces the risk of the words being topically biased toward a particular domain. In addition to the linguistic phenomena, social media provides access to speakers' behavior in a variety of contexts that are not always available from interview studies or participant observation. In Chapter 4 and Chapter 6, I showed that context matters for explaining the form of language that speakers choose, which includes the relative time during which a person mentions a location, and the topical context of a discussion (on-topic versus off-topic). Tracking the same speaker across different contexts reveals how they react to different social constraints that may not arise during a spoken conversation, e.g. the different phases of an ongoing event (Houston et al., 2015). Third, social media provides important insight into a speaker's background that may not be naturally identified in other studies due to problems such as self-reporting bias (Donaldson and Grant-Vallone, 2002). This is particularly important for assessing a speaker's prior language knowledge, in the case of multilingual speakers, and for assessing a speaker's potential prior expectations, in the form of attitudes as well as social status. Both Chapter 3 and Chapter 4 demonstrate how speakers adjust to a given conversation based partly on their relative status within the conversation, whether community membership or overall activity level. While such status obviously plays a role in offline spoken contexts, it may be difficult to map a concept such as community membership onto all participants of a given spoken conversation without having more

context of the participants' previous behavior.

8.1.5 Study design considerations

Any study that investigates sociolinguistic variation in social media should try to leverage the *volume*, *variety*, and *velocity* of social media data (Stieglitz et al., 2018) to address otherwise inaccessible linguistic and social constructs. To that end, I propose several broad study designs for planning computational sociolinguistics studies, which are exemplified by the work in this thesis.

1. One way to assess language variation is to investigate the collective response to *exogenous events* on social media: the content ban in Chapter 3, the crisis events in Chapter 4, and the political referendum in Chapter 6. As an "always-on" data source (Salganik, 2019), social media provides a lens into a wide *variety* of events, as well as events with a high *velocity* that often requires people to adapt their language quickly to changing circumstances. These events help to provide social context to a language variation as it relates to speakers' communicative goals, as in the hashtag variation study where "deeper" hashtags can help speakers avoid the pro-ED content ban.

   Choosing the right type of event is key as many events are irrelevant or highlight forms of variation that are trivial, e.g. the extreme stylistic variation that occurs in response to sports events (*GOOOOOOL* for soccer games) (Brody and Diakopoulos, 2011). This thesis focused on events that occurred over a long period of time (e.g. several weeks for natural disasters in Chapter 4), encompassed a consistent population of speakers, and captured a situation where language variation marks a notable difference in how speakers reacted to the event. Preliminary qualitative analysis of posts made in reaction to an event can help identify whether the event meets such criteria. This step helped justify the study of the independence referendum in Chapter 6, which revealed strong political associations with Catalan

even from the use of particular hashtags (e.g. the Catalan hashtag *#JoVoto* "I vote"). Events that engage speakers repeatedly (e.g. the evolving response to hashtag ban in Chapter 3) can be useful in assessing *within-speaker* variation, which may be less easily understood in spoken conversation where speakers don't have an opportunity to react to different contexts in the same short period of time.

2. The second general approach to study language variation is to focus on *rare* patterns of variation and change: the adoption of nonstandard words (Chapter 5) and the integration of loanwords (Chapter 7). The high *volume* of social media data can unearth phenomena that are rare in spoken contexts, particularly those related to language change which often require multiple generations of people to observe (Poplack, Sankoff, and Miller, 1988; Tagliamonte and D'Arcy, 2007). Rather than relying on a rigid experiment to elicit rare loanword use (Lev-Ari and Peperkamp, 2014), I was able to observe a sufficiently large number of types of loanwords through Twitter discussions (Chapter 7) to characterize consistent variation in loanword structure. Furthermore, the *variety* of data available shows provides space to observe patterns of variation in many different social and linguistic contexts, providing extra insight into possible constraints on rare variables. In Chapter 5, I focused on linguistic dissemination as it related to language change, partly in response to prior studies that had leveraged the variety of *social* contexts available on social media (Altmann, Pierrehumbert, and Motter, 2011) but not the variety of *linguistic* contexts available (Hofmann, Pierrehumbert, and Schütze, 2020).

These "research recipes" are not exhaustive but offer a path forward for sociolinguists who want to expand their horizons into a more computational space. These approaches require sufficient background knowledge about a particular form of variation to ensure that social media will provide the extra context necessary to address open questions, rather than following questions that have already been addressed. For instance,

rather than focusing on network structure as a means of explaining language change (Kershaw, Rowe, and Stacey, 2016; Milroy and Milroy, 1985), a researcher might focus on a different form of social context through which to study change, such as social *identity* which speakers often construct through adoption of new words (Bucholtz and Hall, 2005; Eckert, 2016). Rather than strict rules, this type of research should be guided by intuitions about patterns of variation (e.g. "Do people use descriptor information strategically or is it mostly random?") as well as critical thinking to formulate research questions that best leverage the volume, variety and velocity of social media data ("What social contexts online can provide context to understand descriptor choices, which aren't available in spoken corpora?"). Lastly, a research design should stay flexible in the event that a particular question proves less fruitful than anticipated. In Chapter 7, social attitude (as measured by music sharing) did not explain the integration of loanwords but did explain the integration of native verbs, suggesting an unexpected divergence in how speakers perceive the formality expectations of different word categories.

Determining level of analysis In computational sociolinguistics research, the evaluation should include quantitative hypothesis testing that tests the role of the social or linguistic factors of interest in explaining a regular pattern of language variation. To design the right kind of test, researchers should consider whether the pattern of variation needs to be examined from a low level of analysis (individual speaker decisions) or a high level (collective behavior). A lower-level analysis can provide insight into possible speaker motivations behind language choices, such as audience (Bell, 1984), while a higher-level analysis can reveal emergent behavior within broader social systems (Metcalf, 2004).

In this thesis, I investigate multiple levels of social behavior to address large-scale patterns of variation on social media. For Chapters 4 and 5, I investigate collective trends, including attention "peaks" and dissemination among contexts, to provide global explanations for language variation that would be missed at lower levels. When

investigating collective changes as in the case of nonstandard word adoption, one can discover broad patterns to inform language change theory without necessarily understanding speaker-level decisions. While we do not always know the source from which a speaker adopts a new word on social media (e.g. speaker could adopt *lol* from a website external to Reddit), we can measure aggregate trends in word adoption and abandonment to understand the likely influence of macro-level factors. For Chapters 6 and 7, I investigate speaker-level trends to determine the social value of code-switching for particular speaker groups who could have consistent social motivations for their language choices. For Chapters 3 and 4, I perform analysis at the level of individual utterances (e.g. predicting the "depth" of a hashtag in a post) to determine the role of community or discussion context, in particular *temporal* context, in explaining language choices. In all studies, it is important to scope the conclusions of the study based on the level of analysis. Following the earlier points about study design, the choice of level of analysis should be driven by the specific pattern of language variation and what social media (as opposed to other domains) can offer to address open sociolinguistics questions.

Choosing platform of study When investigating language variation online, it is important to consider the properties of social media platforms that would best suit a particular study. I chose to study *public* discussions on social media for (1) ease of data acquisition, (2) reduced likelihood of ethical considerations, and (3) higher ability for future work to replicate the analysis. Although public data has the added risk of selection bias (people who are willing to share content in public; Hargittai, 2020), the relatively open nature of discussions on most social media platforms provides ample space for a variety of people to be represented.

To address RQ1, I required platforms that hosted discussions which have a dynamic set of participants who needed to participate in a discussion without necessarily knowing their audience (Litt, 2012). In Chapter 3, I chose Instagram as a platform of study due to

its reliance on hashtags to organize dynamic communities of practice (Blight, Ruppel, and Schoenbauer, 2017). For Chapter 4 I chose Twitter due to the platform's popularity among people sharing information with unknown audiences during breaking news events (Kogan, Palen, and Anderson, 2015) I also chose Facebook due to the platform's public groups that have explicit geographic "boundaries" to help contextualize discussion (Bird, Ling, Haynes, et al., 2012). To address RQ2, I chose Reddit due to the platform's frequent language changes (Zhang et al., 2017; Tredici and Fernández, 2018), the diversity of discussion in which words can be adopted, and the multiple levels of social "units" (speakers, threads, communities) to which words could spread. To address RQ3, I required a platform with sufficient multilingual activity (Kim et al., 2014) and which speakers had freedom to express their attitudes regarding different political and cultural groups (Sauter and Bruns, 2015). In Chapters 6 and 7, I therefore chose Twitter to identify diverse populations of multilingual speakers who were likely to leverage the platform to express their attitudes.

8.2 Limitations

Here I address several limitations that apply to all studies in this thesis.

The first limitation is generalizability. The population of people who use social media differs considerably from the general population (Wojcik and Hughes, 2019), which means that the findings of this thesis may not extend to all offline discussions. People who use social media platforms also tend to be younger and more technologically connected (Hargittai, 2020), and the geographic distribution of authors represented on social media often does not align with the offline population (Kariryaa et al., 2018; Pavalanathan and Eisenstein, 2015b). To that end, the findings about social attitude (Chapter 6 and Chapter 7) may be limited by the fact that the online population could have self-selected and therefore have more extreme attitudes than the general population (Tucker et al., 2018). Similarly, for the findings about conversation adaptation

(especially Chapter 4), the speakers are likely in a position where they have free time and the technological capability to respond to an ongoing event, which differentiates them from the overall affected population who likely lacks the means to participate as readily (Soden and Palen, 2018). The studies of this thesis acts should be considered as *situated* analyses that do not necessarily explain social behavior in other areas, given that people have different constraints in online communication as compared to offline life. This point relates especially to the findings about language change, as the time periods studied in this thesis are much shorter than typical sociolinguistics studies which often span several decades in apparent time (D'Arcy and Tagliamonte, 2015; Poplack, Sankoff, and Miller, 1988). More cross-domain research is required to determine the generalizability of findings in language change from social media to spoken language (Kulkarni et al., 2015; Shoemark et al., 2019).

From the data perspective, the thesis relies on high-precision filtering procedures that may reduce the sample size and statistical power of the various studies. In the multilingual studies (Chapter 4, Chapter 6, Chapter 7), posts were filtered based on their confidence score generated by a language ID algorithm (Lui and Baldwin, 2012). The content omitted by this filtering may have included more language mixing which would contribute differently to the results, e.g. loanword integration in the context of code-switching may have different social constraints than in monolingual discourse. For studies in word-level variation (Chapter 5, Chapter 7), I manually curated word lists based on patterns in the data and pre-existing knowledge bases, which may not provide a complete representation of the phenomena under consideration: for the word adoption study, I discarded nonstandard words such as *dope* due to their assumed ambiguity with standard senses. Weakening the data filter on the studies may have provided more statistical power for some of the borderline findings, such as the unclear media consumption finding in Chapter 7, at the risk of generating spurious correlations. As with many computational social science studies, this thesis also has potential for false positive errors in the NLP

pipeline. The Spanish posts identified via `langid` (for Chapter 4, Chapter 6, and Chapter 7) were qualitatively examined but the system may have suffered from systematic false positives that weren't caught, e.g. named entities that triggered a "Spanish" label. These potential errors pose a significant challenge to future research in other forms of variation in language structure, such as African American English syntax (Blodgett, Wei, and O'Connor, 2018), which requires highly accurate dependency parsing. Similarly to other work in computational social science (Zamith and Lewis, 2015), future research in computational sociolinguistics should consider the trade-offs of automatic and manual coding of linguistic phenomena in terms of construct validity.

While this thesis relies almost entirely on empirical quantitative analysis, this thesis largely presents correlational and *not causal* (Salganik, 2019) evidence in favor of the hypotheses tested. With the exception of Chapter 5 which leverages a form of causal inference, all analysis relies on regression and statistical hypothesis testing. In most studies, I focus on the relative influence of factors on language variation rather than determining the true causal impact of any one factor, which still fulfills the overall goal of explanation as opposed to prediction (Hofman, Sharma, and Watts, 2017). However, in some analyses it may be the case that the effects observed only partly explain a language pattern and are best explained by latent factors that could not be controlled. In Chapter 6 and Chapter 7, we have no guarantee that political stance and media consumption are not affected by an underlying factor that is the actual cause of the language patterns observed, such as underlying personality traits (Park et al., 2015). Unlike typical experimental studies (Lev-Ari and Peperkamp, 2014), the studies in this thesis are observational and do not guarantee that researchers will be able to replicate exactly the effects observed in a controlled setting. Some of the studies try to control for confounds with speaker-level and word-level effects (Chapter 4, Chapter 7), which may reduce the chance of spurious effects from phenomena such as population imbalance. Addressing (approximately) causal relationships could require a wider pool of speakers with more diversity in their

traits to allow for strategies such as approximate matching between "treated" and "control" groups of speakers (Chandrasekharan et al., 2017). Causal modeling strategies do not easily extend to the studies of this thesis because the linguistic phenomena considered are quite rare, which naturally reduces the visible speaker population. For instance, assessing the relationship between time and descriptor use in Chapter 4 would require a large pool of location names to match on location size and salience to the discussion, which is difficult because there is a limited pool of names available for analysis. However, studying causality would clarify the findings for studies that occurred in the context of an external event (content banning, crisis events, political discussions), because it may be the case, similar to Pavalanathan and Eisenstein (2016), that the event itself tends to change the linguistic behavior of participants more than the participants' own audience awareness or attitudes alone.

In terms of more general methods, a more thorough approach would involve qualitative investigation to address gaps in the findings, such as the unusual difference in integration rates between loanwords and native verbs in Chapter 7. Such qualitative methods are common to discourse analysis (Ferrara and Bell, 1995; Maíz-Arévalo and García-Gómez, 2013) and include close reading of texts to infer common *intentions* among speakers in their choice of variants based on the conversation context. This thesis focused on quantitative analysis in part to ensure replicability in future work, but more qualitative evaluation, such as the inspection of reply-tweets in Chapter 6, would improve the studies' contributions to the *why* of language variation. More qualitative analysis would also help with studies where quantitative methods fall short of completely addressing the research questions. For the studies in social attitudes, even a completely causal model (e.g. comparing loanword integration before/after sharing media) would not help assess the *social value* of the speakers' expressed attitudes (Bucholtz, 1999), which qualitative analysis would help address (e.g. whether high-SLA media speakers tend to express opinions about one culture versus another even in posts that aren't about music).

From a social theory perspective, this thesis largely ignores network structure as a factor in language variation and change (Milroy and Milroy, 1985) in favor of audience (RQ1), context spread (RQ2), and attitudes (RQ3). A more network-centric perspective would help unravel remaining questions regarding the construction of "community": in Chapter 3, how readily are newcomers integrated into the pro-ED community structure, and does this explain how quickly they abandon the more extreme variants? For the "natural experiment" studies that relied on reactions to specific events (Chapters 3, 4 and 6), a network perspective would reveal whether the collective response to an ongoing event represented a coherent community, with corresponding language norms (e.g. a community of strong ties may not need descriptors). Furthermore, adding network information for individual speakers would help clarify the role of social *relationships* in explaining language variation (Danescu-Niculescu-Mizil, Gamon, and Dumais, 2011; Maíz-Arévalo and García-Gómez, 2013). For instance, would a speaker tend to add descriptors (Chapter 4) or switch to the majority language (Chapter 6) if their listener was not previously connected to them, to signal a more polite relationship? With respect to broader change processes such as adoption of new words and loanwords, it is important to consider whether particularly sparse areas of the network (cf. weak ties; Granovetter, 1973) are more responsible for accelerating change, or if some of the new words originated from dense "trend-setter" communities (Grieve, Nini, and Guo, 2018; Young, 2011) (e.g. *stalkear* emerging from cliques of younger people).

8.3 Implications for future work

The work in this thesis will inform future research in computational sociolinguistics and social computing.

8.3.1 Practical applications

From a systems perspective, understanding patterns of variation in structure can have implications for canonical NLP tasks, such as understanding biases in performance across social groups for parsing or POS tagging (Garimella et al., 2019; Johannsen, Hovy, and Søgaard, 2015). If a particular dialect or speech community has significant variation in language structure, this may require collecting and re-annotating new data to guarantee fair performance on downstream tasks (Blodgett, Wei, and O'Connor, 2018). In a more direct application, understanding regular patterns of variation can help engineers build systems that actively adapt to a speaker's language use, such as text prediction that can accurately model code-switching and loanword use (Solorio and Liu, 2008). Building more socially aware systems will require further testing to determine the situations in which multilingual speakers expect code-switched compatibility, as some speaker groups such as immigrants may expect code-switching more often to navigate their social life (Papalexakis, Nguyen, and Doğruöz, 2014). As NLP systems such as automatic translation become ubiquitous on social media platforms (Duarte, Llanso, and Loup, 2018), it will be important to make room for multilingual speech to be accurately modeled. Outside of prediction, socially-aware NLP systems can provide writers with tools for self-reflection when preparing writing with a specific social motivation. Current writing interventions often focus on mitigating negative behavior (Chandrasekharan et al., 2019), but future systems should embrace a wide range of social motivations to help people adjust to their audience even in positive settings. When writing messages on social media (Frankenberg-Garcia et al., 2019), an author could provide a writing tool with the intended audience, which would allow the tool to make more relevant suggestions (e.g. "include more context for *San Juan*, the readers may not know about it"). In general, NLP developers should consider how sociolinguistic theory can inform speakers' social motivations for communication and therefore what speakers expect from systems that claim to improve communication (Blodgett et al., 2020).

Future work in studying language variation can also inform broader questions about the design of social computing systems, through systematic study of computer mediated communication. For one, this thesis suggests that people prepare for their audience not just with typical style cues such as function words (Danescu-Niculescu-Mizil, Gamon, and Dumais, 2011), but also with variation in their syntax. Researchers in computational social science have recently begun to investigate the role of audience in communication such as emails (Zhang et al., 2020) and social media posts (Kaur, Lampe, and Lasecki, 2020), with the goal of developing interventions to help writers. Comparing the types of linguistic adaptation across writing domains (email versus social media versus press releases) would reveal the nature of audience expectations in different scenarios, thereby providing a more fine-grained view into how people prepare for possible feedback from their listeners. Furthermore, studying speakers' adaptation to their conversation expectations at scale can provide another tool for those monitoring public reactions to events on social media (Varga et al., 2013). If a sub-group in a community provides less description for their audience about a topic of discussion, it may indicate a shared common ground with respect to that topic and therefore a stronger in-group orientation. This is particularly important for stakeholders such as government crisis responders who may need to know about real-time public awareness of unfolding events to be able to plan their responses accordingly (Soden and Palen, 2018).

With respect to social computing methods, this thesis presents an array of linguistic phenomena to measure that can complement more typical content analysis techniques, such as lexicons or topic models. This perspective can particularly help in general cases of long-term change at the individual level (e.g. reactions to platform changes), where changes in individual word counts may be driven more by ephemeral influences like news events. A person who consistently shows lower linguistic dissemination (Chapter 5) in a particular word over time may be implicitly preparing to abandon the word due to lower commitment over time, even if the aggregate frequency remains the same. The

communicative intent in the language choices studied in thesis will likely generalize to a wide variety of social contexts, as well. The respelled words exemplified by Chapter 3 are not just used to evade content bans (Chancellor et al., 2016), they can also be used in cases of general aversion to a concept due to disgust (McCulloch, 2019), or in cases where a new group of people or bots attempts to "claim" new discussion territory by modifying an existing popular hashtag. Lastly, researchers can readily modify the language choices proposed in this thesis to accommodate different analysis goals. The short-range dependency arcs used by Chapter 4 to detect descriptors can be repurposed to match different kinds of descriptive information such as adjectives (*the Canadian Justin Trudeau*) and coordinating phrases (*Atlanta, or "Hotlanta" according to some*), which can be useful in understanding more indirect audience design.

8.3.2 Theoretical considerations

Research in computational sociolinguistics has consistently focused on lexical variation (Bamman, Eisenstein, and Schnoebelen, 2014; Pavalanathan and Eisenstein, 2015a), due to the relative ease of counting words and high interpretability in results. By investigating word structure and phrase structure, this thesis "pushes the envelope" of variation (Aaron, 2010) and proposes new ways forward in the investigation of language variation. As a variable of study, language structure is generative, detectable with surface-level NLP to varying degrees (e.g. simple syntax is more feasible than complex syntax) and applicable to situations with sparse data. Focusing on structure could also be useful in the study of languages other than English (Stanford, 2016), which has dominated sociolinguistic inquiry due to high availability of data but represents a limited distribution of linguistic traits (e.g. simple morphology). Future work into code-switching should consider the wide range of social attributes available in online discussions, including not just attitude expression but also affiliation with context-specific social systems such as tribal membership (Ndubuisi-Obi, Ghosh, and Jurgens, 2019).

In addition, studying long-term language change on social media can reveal the linguistic restrictions on change that may be more important than social factors alone. Designing metrics for different aspects of linguistic context, e.g. semantic scope (Ryskina et al., 2020), can help address what exactly about a word's use makes someone likely to adopt the word, similar to how more fine-grained social metrics can identify the network patterns that lead to adoption (Goel et al., 2016; Leskovec, Backstrom, and Kleinberg, 2009). Semantic representations of text such as word embeddings will help define more generic notions of linguistic utility, including whether a word like "insanely" has become semantically *bleached* enough to be useful as an intensifier in any adjective context (Luo, Jurafsky, and Levin, 2019). In general, deep learning methods can help provide a more accurate representation of linguistic structures that underlie language change, such as contextual word representations (Giulianelli, Del Tredici, and Fernández, 2020) that reveal the changes in sense distributions of a word at different points in time. From a theory perspective, the distinction between grammatical (Ito and Tagliamonte, 2003) and topical (Karjus et al., 2018) context would be particularly useful to study, e.g. whether a word succeeds because of its grammatical utility or because of its relevance to a variety of topics. Studying the change in such contextual metrics over a long period of time can also reveal different patterns of "entrenchment" in different speech communities (Chesley and Baayen, 2010). Some words may disseminate broadly while others settle into smaller niches, such as the different trajectories of computer-related words like *reboot* (wide range of contexts) and *keyboard* (restricted contexts).

This thesis focused more on the production side of language variation rather than the reception or *perception* of variation (Preston, 2013), which is equally important: can we quantify how often people actively notice significant language variation and changes in progress in online discussions? Anecdotally I noticed several examples of meta-commentary on language use in Chapter 5 (*who says* cringey *anymore*) and Chapter 7 (googlear *sounds like something my grandpa would say*). I expect that more

visible forms of variation, such as word use and language structure, would be more readily perceived by the general public and considered change from above (Labov, 1973) as compared to more subtle forms, such as semantic change, which would be considered change from below. Although such perceptions can be captured by explicit surveys (Del Tredici, Fernández, and Boleda, 2019), a more scalable approach would entail automatically finding explicit reactions to language variation and change in social media. Such reactions may manifest through parody behavior (dialect joke accounts; Tatman, 2016) or through humor in response to use of unusual linguistic forms (*I can't believe you used the word* dope *unironically*; Childs and Mallinson, 2006).

From a social computing perspective, the thesis presents several theoretical considerations for the study of online community dynamics (Kraut and Resnick, 2012) with respect to member status and norm development. Future work should consider how different community members contribute to local versus global changes in communication norms. A new member in a particular sub-community may feel less willing to contribute to a change in progress in the community but more willing to participate in global change across all communities, due to the difference in social value between the changes (e.g. the globally appealing *lol* versus the community-specific laughter *kek*). Furthermore, the studies in this thesis consider changes in progress at different stages (e.g. the growth stage and the decline stage), and future work should compare the cycle of change across communities (Zhang et al., 2017) to determine the cycle's consistency in different social systems. Communities that have strong language norms (August et al., 2020; Chandrasekharan et al., 2018), such as `r/science`, may implicitly discourage the adoption of new norms and have a more "static" life cycle than more welcoming communities. Further work should look into the formal norms that communities establish for their members (e.g. neutral language use; Pavalanathan, Han, and Eisenstein, 2018) and determine to what degree these prescriptive norms affect the natural cycles of variation and change that occur in online communities (Tan, 2018).

Appendices

APPENDIX A

COLLECTIVE ATTENTION

A.1 Detecting author social status

In the context of event-based public discussions, it is worth considering whether a post author is (1) local and (2) an organization. An author who is *local* (more committed) to the event's region will already be aware of the locations under discussion (Kogan, Palen, and Anderson, 2015) and will be less likely to use context than an author who is unfamiliar with the region's locations. Organizations such as government agencies are often responsible for disseminating official information to help crisis responders and effectively organize aid (Houston et al., 2015). An author who represents an official organization may want to minimize uncertainty in their messages and use more context than an author who does not represent an organization, i.e. a citizen observer.

I determined author local status and organization status using a sample of metadata available from archived tweets corresponding to the time periods of interest (covering $\sim 20\%$ of all authors in the data). Following prior work in geolocation (e.g. Kariryaa et al., 2018), I approximate the local status of an author posting about an event based on whether the author's self-reported profile location mentions a relevant city or state in the event's affected region (e.g. for Hurricane Maria, a local author would mention *Puerto Rico* or *PR* in their location field).

Organizations are difficult to identify automatically, because there is no single indicator of organization status in a Twitter user's profile information. To determine whether an author counts as an organization, I apply the pre-trained organization classifier from Wood-Doughty, Mahajan, and Dredze, 2018[1] to the author's metadata, including

[1] Accessed 7/2019: `https://bitbucket.org/mdredze/demographer/src/peoples2018/`.

Table A.1: Regression results for Facebook data in RQ1, using `spacy` parses to detect descriptor phrases. * indicates $p < 0.05$, otherwise $p > 0.05$.

Factor	Variable	Estimate	S.E.
Intercept		-2.08	34.4
Importance	Prior location mentions	-0.042	8.31
Author	Author in-group posts	-0.172	0.549
Audience	Local location	-0.607*	0.116
	Group size	0.106	40.3
Deviance			469
Accuracy			74.3%

name, description, and social attributes.

For both local and organization status, I find reasonable precision with respect to a small subset of hand-labeled authors from my data.[2]

A.2 Robustness check for descriptor extraction

As mentioned in § 4.2.3, I used the SyntaxNet parser to extract descriptor phrases from the Facebook data due to the parser's better performance on longer sentences. To verify the consistency of results across parsers, I re-parsed the Facebook data with the `spacy` parser used for the Twitter and repeated the regression to predict descriptor use from the explanatory factors, i.e. RQ1. The effect sizes and significance remained the same in the regression on spacy parses, shown in Table A.1 (cf. "RQ1 (Facebook)" column in Table 4.5).

[2]I annotated 500 accounts as organizations and locals, based on available metadata, and compared these labels to those produced by the local proxy and organization classifier. The local proxy achieved precision of 87% and recall of 58%, and the organization classifier achieved precision of 87% and recall of 54%.

APPENDIX B

LOANWORD INTEGRATION

B.1 All integrated and light verb pairs

To assist replication, I list all pairs of integrated and light verbs for loanwords and native verbs used in the loanword integration study (§ 7.2). I list them in alphabetical order (by integrated verb) in the format:

loanword/translation: *integrated verb* ; *light verb(s)*

Loanwords

- **access**: *accesar* ; *hacer/tener acces*
- **aim**: *aimear* ; *hacer/tener aim*
- **alert**: *alertear* ; *hacer alert*
- **audit**: *auditar* ; *hacer (un) audit*
- **ban**: *banear* ; *hacer un ban*
- **bang**: *bangear* ; *hacer bang*
- **bash**: *bashear* ; *hacer/dar bash*
- **block**: *blockear* ; *hacer/dar (un) block*
- **boycott**: *boicotear* ; *hacer (un) boicot*
- **box**: *boxear* ; *hacer (el) box/boxing*
- **bully**: *bulear* ; *hacer/ser (el) bully*
- **bust**: *bustear* ; *hacer (el) bust*
- **cast**: *castear* ; *hacer cast/casting*
- **change**: *changear* ; *hacer change*
- **chat**: *chatear* ; *hacer chat*
- **check**: *chequear* ; *hacer un cheque*
- **shoot**: *chutar* ; *hacer/tomar el shot*
- **combat**: *combatear* ; *hacer (el) combat*
- **connect**: *conectar* ; *hacer (un) conexión*
- **crack**: *crackear* ; *hacer crack*
- **customize**: *customizar* ; *hacer custom/customized*
- **default**: *defaultear* ; *hacer default*
- **delete**: *deletear* ; *hacer/poner delete*
- **DM**: *dmear* ; *mandar/enviar/poner un dm*
- **dope**: *dopar* ; *hacer doping*
- **downvote**: *downvotear* ; *poner/dar (un) downvote*
- **draft**: *draftear* ; *hacer/tener draft*
- **drain**: *drenar* ; *hacer (el) dren*

- **dribble**: *driblar* ; *hacer (el) drible*
- **encrypt**: *encriptar* ; *hacer/ser encript*
- **smash**: *esmachar* ; *hacer smash*
- **sniff**: *esnifar* ; *hacer sniff*
- **standard**: *estándar* ; *hacer (un) standard*
- **exit**: *exitear* ; *hacer exit*
- **export**: *exportear* ; *hacer export*
- **externalize**: *externalizar* ; *hacer external*
- **fangirl**: *fangirlear* ; *hacer/ser fangirl*
- **film**: *filmar* ; *tomar (un) film*
- **flash**: *flashear* ; *hacer (un) flash*
- **flex**: *flexear* ; *hacer (un) flex*
- **flip**: *flipar* ; *hacer flip*
- **flirt**: *flirtear* ; *hacer flirt*
- **focus**: *focalizar* ; *hacer focus*
- **format**: *formatear* ; *hacer/dar (el) formato*
- **form**: *formear* ; *hacer form*
- **freak**: *friquear* ; *estar freaked*
- **freeze**: *frizar* ; *hacer freeze*
- **fund**: *fundear* ; *dar/hacer fund/funding*
- **gentrify**: *gentrificar* ; *hacer/tener gentrificación*
- **ghost**: *gostear* ; *hacer gost/ghost*
- **google**: *googlear* ; *buscar en google*
- **hack**: *hackear* ; *hacer hack*
- **hail**: *hailear* ; *hacer hail*
- **hang**: *hanguear* ; *hacer hang*
- **harm**: *harmear* ; *hacer harm*
- **hypnosis**: *hipnotizar* ; *hacer hipnosis*
- **host**: *hostear* ; *hacer host*
- **hype**: *hypear* ; *hacer hype*
- **intercept**: *interceptear* ; *hacer/tirar interception*
- **hang**: *janguear* ; *hacer hang (out)*
- **lag**: *lagear* ; *hacer (un) lag*
- **like**: *likear* ; *dar/poner (un) like*
- **limit**: *limitear* ; *hacer (un) limit*
- **lynch**: *linchar* ; *hacer lynch*
- **link**: *linkear* ; *dar/poner (un) link*
- **love**: *lovear* ; *hacer love*
- **look**: *luquear* ; *dar/hacer (un) look*
- **make**: *makear* ; *hacer make*
- **melt**: *meltear* ; *hacer melt*
- **mope**: *mopear* ; *hacer mope*
- **nag**: *nagear* ; *hacer nag*
- **knock**: *noquear* ; *dar/hacer (un) knockout*
- **pack**: *packear* ; *hacer pack*
- **pan**: *panear* ; *hacer/dar (un) panorama*
- **panic**: *paniquear* ; *tener panic*
- **park**: *parquear* ; *hacer parking*
- **perform**: *performar* ; *hacer (un) performance*
- **pitch**: *pichear* ; *hacer (un) pitch*

- **pin**: *pinear* ; *hacer pin*
- **PM**: *pmear* ; *enviar/mandar (un) pm*
- **punch**: *ponchar* ; *hacer un punch*
- **post**: *postear* ; *dar/poner (un) post*
- **posterize**: *posterizar* ; *hacer poster*
- **print**: *printear* ; *hacer print*
- **protest**: *protestear* ; *hacer (un) protest*
- **push**: *puchar* ; *hacer un push*
- **pump**: *pumpear* ; *hacer pump(s)*
- **quote**: *quotear* ; *hacer quote*
- **rank**: *rankear* ; *hacer rank*
- **rant**: *rantear* ; *hacer (un) rant*
- **rape**: *rapear* ; *hacer (un) rape*
- **record**: *recordear* ; *hacer (un) recording*
- **remaster**: *remasterizar* ; *hacer remastered*
- **render**: *renderizar* ; *hacer render(ed)*
- **rent**: *rentear* ; *hacer rental/renting*
- **report**: *reportear* ; *hacer (un) report*
- **reset**: *resetear* ; *hacer reset*
- **respect**: *respectear* ; *hacer respect*
- **ring**: *ringear* ; *hacer ring*
- **rock**: *rockear* ; *hacer rock*
- **roll**: *rollear* ; *hacer roll*
- **sample**: *samplear* ; *hacer (un) sample*
- **selfie**: *selfiar* ; *tomar (un) selfie*
- **sext**: *sextear* ; *dar/mandar un sext*
- **ship**: *shippear* ; *hacer ship*
- **shitpost**: *shitpostear* ; *hacer/poner un shitpost*
- **shock**: *shockear* ; *hacer shock*
- **sign-in**: *signear* ; *hacer sign-in*
- **stalk**: *stalkear* ; *actuar como un stalker*
- **strike**: *strikear* ; *hacer/dar un strike*
- **surf**: *surfear* ; *hacer surf*
- **tackle**: *taclear* ; *hacer tackle*
- **text**: *textear* ; *mandar/enviar un text*
- **tick**: *ticar* ; *hacer (un) tick*
- **torment**: *tormentear* ; *hacer torment*
- **touch**: *touchear* ; *hacer (un) touch*
- **transport**: *transportear* ; *hacer transport*
- **travel**: *travelear* ; *hacer travel*
- **troll**: *trolear* ; *actuar como un trol*
- **tweet**: *tweetear* ; *poner/enviar/hacer (un) tweet*
- **twerk**: *twerkear* ; *hacer twerk*
- **upvote**: *upvotear* ; *dar (un) upvote*
- **vape**: *vapear* ; *hacer/tomar vape/vaping*
- **zap**: *zapear* ; *hacer zap/zapping*

Native verbs

- **admire**: *admirar* ; *tener admiración*
- **befriend**: *amistar* ; *tener amistad*
- **encourage**: *animar* ; *subir el ánimo*
- **note**: *anotar* ; *tomar nota*
- **land**: *aterrizar* ; *hacer un aterrizaje*
- **joke**: *bromear* ; *hacer bromas*
- **mock**: *burlarse* ; *hacer burla*
- **punish**: *castigar* ; *poner un castigo*
- **buy**: *comprar* ; *hacer la compra*
- **copy**: *copiar* ; *hacer una copia*
- **tickle**: *cosquillar* ; *hacer cosquillas*
- **blame**: *culpar* ; *echar la culpa*
- **damage**: *dañar* ; *hacer daño*
- **decide**: *decidir* ; *tomar decisiones*
- **apologize**: *disculparse* ; *pedir disculpas*
- **shower**: *ducharse* ; *darse una ducha*
- **question**: *dudar* ; *poner en duda*
- **exemplify**: *ejemplificar* ; *poner un ejemplo*
- **estimate**: *estimar* ; *tener estima*
- **explain**: *explicar* ; *dar una explicación*
- **finish**: *finalizar* ; *poner fin*
- **photograph**: *fotografiar* ; *tomar fotos*
- **escape**: *fugarse* ; *darse a la fuga*
- **mention**: *mencionar* ; *hacer mención*
- **look at**: *mirar* ; *echar una mirada*
- **penalize**: *multar* ; *poner una multa*
- **negotiate**: *negociar* ; *hacer negocios*
- **originate**: *originar* ; *dar origen*
- **participate**: *participar* ; *tomar parte*
- **walk**: *pasear* ; *dar un paseo*
- **step**: *pisar* ; *poner el pie*
- **value**: *preciar* ; *poner precio*
- **ask**: *preguntar* ; *hacer (una) pregunta*
- **anticipate**: *prever* ; *tener previsto*
- **test**: *probar* ; *poner a prueba*
- **recommend**: *recomendar* ; *hacer recomendación*
- **write**: *redactar* ; *hacer una redacción*
- **cure**: *remediar* ; *poner remedio*
- **breathe**: *respirar* ; *dar un respiro*
- **jump**: *saltar* ; *dar un salto*
- **nap**: *sestear* ; *echar una siesta*
- **dream**: *soñar* ; *tener un sueño*
- **end**: *terminar* ; *poner término*
- **use**: *usar* ; *hacer uso*
- **travel**: *viajar* ; *hacer un viaje*
- **see**: *vistar* ; *echar un vistazo*
- **fly**: *volar* ; *tomar un vuelo*

REFERENCES

[1] Jessi Elana Aaron. “Pushing the envelope: Looking beyond the variable context”. In: *Language variation and change* 22.1 (2010), pp. 1–36.

[2] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. “Privacy and human behavior in the age of information”. In: *Science* 347.6221 (2015), pp. 509–514.

[3] Eric K Acton and Christopher Potts. “That straight talk: Sarah Palin and the sociolinguistics of demonstratives”. In: *Journal of Sociolinguistics* 18.1 (2014), pp. 3–31.

[4] Tim Althoff and Jure Leskovec. “Donor retention in online crowdfunding communities: A case study of donorschoose.org”. In: *WWW*. 2015, pp. 34–44.

[5] Eduardo Altmann, Janet Pierrehumbert, and Adilson Motter. “Niche as a determinant of word fate in online groups”. In: *PLoS ONE* 6.5 (2011), pp. 1–12.

[6] Roberta Amato, Lucas Lacasa, Albert Díaz-Guilera, and Andrea Baronchelli. “The dynamics of norm change in the cultural evolution of language”. In: *Proceedings of the National Academy of Sciences* 115.33 (2018), pp. 8260–8265.

[7] Lieselotte Anderwald. “Measuring the success of prescriptivism: quantitative grammaticography, corpus linguistics and the progressive passive”. In: *English Language & Linguistics* 18.1 (2014), pp. 1–21.

[8] Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. “Globally Normalized Transition-Based Neural Networks”. In: *ACL*. 2016, pp. 2442–2452.

[9] Jannis Androutsopoulos. “Language change and digital media: a review of conceptions and evidence”. In: *Standard Languages and Language Standards in a Changing Europe*. Ed. by Kristiansen Tore and Nikolas Coupland. Oslo: Novus Press, 2011, pp. 145–159.

[10] Jannis Androutsopoulos. “Language choice and code-switching in German-based diasporic web forums”. In: *The multilingual Internet: Language, culture, and communication online* (2007), pp. 340–361.

[11] Jannis Androutsopoulos. *Mediatization and sociolinguistic change*. Vol. 36. Walter de Gruyter, 2014.

[12] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. “Identification of Causal Effects Using Instrumental Variables”. In: *Source Journal of the American Statistical Association* 91.434 (1996), pp. 444–455.

[13] Peter Auer. *Code-switching in conversation: Language, interaction and identity*. London, UK: Routledge, 2013.

[14] Tal August, Dallas Card, Gary Hsieh, Noah A Smith, and Katharina Reinecke. “Explain like I am a Scientist: The Linguistic Barriers of Entry to r/science”. In: *CHI*. 2020, pp. 1–12.

[15] Molly Babel. “Dialect divergence and convergence in New Zealand English”. In: *Language in Society* (2010), pp. 437–456.

[16] Richard W Bailey. “Scots and Scotticisms: language and ideology”. In: *Studies in Scottish Literature* 26.1 (1991), p. 7.

[17] Eytan Bakshy, Jake Hofman, Winter Mason, and Duncan Watts. “Everyone’s an influencer: quantifying influence on Twitter”. In: *WSDM*. 2011, pp. 65–74.

[18] David Bamman, Chris Dyer, and Noah A Smith. “Distributed representations of geographically situated language”. In: *ACL*. 2014, pp. 828–834.

[19] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. “Gender identity and lexical variation in social media”. In: *Journal of Sociolinguistics* 18.2 (2014), pp. 135–160.

[20] George A Barnett, Omar Souki Oliveira, and J David Johnson. “Multilingual language use and television exposure and preferences: The case of Belize”. In: *Communication Quarterly* 37.4 (1989), pp. 248–261.

[21] Allan Bell. “Audience accommodation in the mass media”. In: *Contexts of accommodation: Developments in applied sociolinguistics* (1991), pp. 69–102.

[22] Allan Bell. “Language style as audience design”. In: *Language in society* 13.2 (1984), pp. 145–204.

[23] Allan Bell. *The language of news media*. Blackwell Oxford, 1991.

[24] Michael Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Greg Vargas. “4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community”. In: *ICWSM* (2011), pp. 50–57.

[25] Douglas Biber and Susan Conrad. *Register, genre, and style*. Cambridge University Press, 2019.

[26] Deanne Bird, Megan Ling, Katharine Haynes, et al. “Flooding Facebook-the use of social media during the Queensland and Victorian floods”. In: *Australian Journal of Emergency Management, The* 27.1 (2012), p. 27.

[27] Renée Blake. “Defining the envelope of linguistic variation: The case of “don’t count” forms in the copula analysis of African American Vernacular English”. In: *Language Variation and Change* 9.1 (1997), pp. 57–79.

[28] Katherine Blashki and Sophie Nichol. “Game geek’s goss: linguistic creativity in young males within an online university forum”. In: *Australian Journal of Emerging Technologies and Society* 3.2 (2005), pp. 71–80.

[29] Michael G Blight, Erin K Ruppel, and Kelsea V Schoenbauer. “Sense of community on Twitter and Instagram: Exploring the roles of motives and parasocial relationships”. In: *Cyberpsychology, Behavior, and Social Networking* 20.5 (2017), pp. 314–319.

[30] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. “Language (Technology) is Power: A Critical Survey of “Bias” in NLP”. In: *ACL*. 2020.

[31] Su Lin Blodgett, Lisa Green, and Brendan O’Connor. “Demographic Dialectal Variation in Social Media: A Case Study of African-American English”. In: *EMNLP* (2016), pp. 1119–1130.

[32] Su Lin Blodgett, Johnny Wei, and Brendan O’Connor. “Twitter universal dependency parsing for African-American and mainstream American English”. In: *ACL*. 2018, pp. 1415–1425.

[33] Jan-Petter Blom, John J Gumperz, et al. “Social meaning in linguistic structure: Code-switching in Norway”. In: *The bilingualism reader* (2000), pp. 111–136.

[34] Jan Blommaert. “The long language-ideological debate in Belgium”. In: *Journal of Multicultural Discourses* 6.3 (2011), pp. 241–256.

[35] Richard A Blythe and William Croft. “S-curves and the mechanisms of propagation in language change”. In: *Language* (2012), pp. 269–304.

[36] danah boyd and Kate Crawford. “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon”. In: *Information, communication & society* 15.5 (2012), pp. 662–679.

[37] danah boyd and Nicole Ellison. “Social network sites: Definition, history, and scholarship”. In: *Journal of computer-mediated communication* 13.1 (2007), pp. 210–230.

[38] William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. “Emotion shapes the diffusion of moralized content in social networks”. In: *Proceedings of the National Academy of Sciences* 114.28 (2017), pp. 7313–7318.

[39] Samuel Brody and Nicholas Diakopoulos. “Cooooooooooooooollllllllllllll!!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs”. In: *EMNLP*. 2011, pp. 562–570.

[40] Penelope Brown, Stephen C Levinson, and Stephen C Levinson. *Politeness: Some universals in language usage*. Vol. 4. Cambridge University Press, 1987.

[41] Susan L Bryant, Andrea Forte, and Amy Bruckman. “Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia”. In: *GROUP*. 2005, pp. 1–10.

[42] Bethany Bryson. ““Anything but heavy metal”: Symbolic exclusion and musical dislikes”. In: *American sociological review* (1996), pp. 884–899.

[43] Taina Bucher and Anne Helmond. “The affordances of social media platforms”. In: *The SAGE handbook of social media* (2017), pp. 233–253.

[44] Mary Bucholtz. ““Why be normal?”: Language and identity practices in a community of nerd girls”. In: *Language in society* 28.2 (1999), pp. 203–223.

[45] Mary Bucholtz and Kira Hall. “Identity and interaction: A sociocultural linguistic approach”. In: *Discourse studies* 7.4-5 (2005), pp. 585–614.

[46] Louisa Buckingham. “Light verb constructions in Latin American newspapers: Creative variants and coinages”. In: *Spanish in Context* 10.1 (2013), pp. 114–135.

[47] Joan L. Bybee. “From Usage to Grammar: The Mind’s Response to Repetition”. In: *Language* 82.4 (2006), pp. 711–733.

[48] Rena Torres Cacoullos and Jessi Elana Aaron. “Bare English-origin nouns in Spanish: Rates, constraints, and discourse functions”. In: *Language Variation and Change* 15.3 (2003), pp. 289–328.

[49] Andreea Simona Calude, Steven Miller, and Mark Pagel. “Modelling loanword success - A sociolinguistic quantitative study of Māori loanwords in New Zealand English”. In: *Corpus Linguistics and Linguistic Theory* 16.1 (2017), pp. 29–66.

[50] Jack K Chambers and Peter Trudgill. *Dialectology*. Cambridge University Press, 1998.

[51] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. "#thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities". In: *CSCW*. 2016, pp. 1201–1213.

[52] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. "Crossmod: A cross-community learning-based system to assist reddit moderators". In: *Proceedings of the ACM on Human-Computer Interaction* 3 (2019), pp. 1–30.

[53] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech". In: *Proceedings of the ACM on Human-Computer Interaction* 1 (2017), pp. 1–22.

[54] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. "The internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro Scales". In: *Proceedings of the ACM on Human-Computer Interaction* 2 (2018), pp. 1–25.

[55] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. "Antisocial Behavior in Online Discussion Communities". In: *ICWSM*. 2015.

[56] Paula Chesley. "Lexical borrowings in French: Anglicisms as a separate phenomenon". In: *Journal of French Language Studies* 20.3 (2010), pp. 231–251.

[57] Paula Chesley and Harald Baayen. "Predicting new words from newer words: Lexical borrowings in French". In: *Linguistics* 48.6 (2010), pp. 1343–1374.

[58] Becky Childs and Christine Mallinson. "The significance of lexical items in the construction of ethnolinguistic identity: A case study of adolescent spoken and online language". In: *American Speech* 81.1 (2006), pp. 3–30.

[59] Susan Chinn. "A simple method for converting an odds ratio to effect size for use in meta-analysis". In: *Statistics in medicine* 19.22 (2000), pp. 3127–3131.

[60] Noam Chomsky. *Barriers*. Vol. 13. MIT press, 1986.

[61] M. Sidury Christiansen. ""A ondi queras": Ranchero identity construction by U.S. born Mexicans on Facebook". In: *Journal of Sociolinguistics* 19.5 (2015), pp. 688–702.

[62] Steven Coats. “Variation of New German Verbal Anglicisms in a Social Media Corpus”. In: *Proceedings of the 6th conference on CMC and social media corpora for the humanities*. 2018.

[63] Paul Cook. “Exploiting linguistic knowledge to infer properties of neologisms”. PhD thesis. University of Toronto, 2010.

[64] Paul Cook. “Using social media to find English lexical blends”. In: *Proceedings of the 15th EURALEX International Congress*. 2012, pp. 846–854.

[65] Nikolas Coupland. “Dialect stylization in radio talk”. In: *Language in society* (2001), pp. 345–375.

[66] David Cox. “Regression models and life tables”. In: *Journal of the Royal Statistical Society* 34 (1972), pp. 187–220.

[67] Kathryn Crameri. *Language, the novelist and national identity in post-Franco Catalonia*. Oxford, UK: Routledge, 2017.

[68] Yves Croissant and Giovanni Millo. “Panel data econometrics in R: The plm package”. In: *Journal of Statistical Software* 27.2 (2008), pp. 1–43.

[69] Lane Crothers. *Globalization and American popular culture*. Rowman & Littlefield, 2017.

[70] Patricia Cukor-Avila. “Revisiting the observer’s paradox”. In: *American Speech* 75.3 (2000), pp. 253–254.

[71] Tiago Cunha, David Jurgens, Chenhao Tan, and Daniel Romero. “Are all successful communities alike? characterizing and predicting the success of online communities”. In: *WWW*. 2019, pp. 318–328.

[72] Cecilia A Cutler. “Yorkville crossing: White teens, hip hop and African American English”. In: *Journal of sociolinguistics* 3.4 (1999), pp. 428–442.

[73] Jennifer Dailey-O’Cain. “The sociolinguistic distribution of and attitudes toward focuser like and quotative like”. In: *Journal of Sociolinguistics* 4.1 (2000), pp. 60–80.

[74] Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. “You had me at hello: How phrasing affects memorability”. In: *ACL*. 2012, pp. 892–901.

[75] Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. “Mark my words! Linguistic style accommodation in social media”. In: *WWW*. 2011, pp. 745–754.

[76] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. “No country for old members: User lifecycle and linguistic change in online communities”. In: *WWW*. 2013, pp. 307–318.

[77] Alexandra D’Arcy and Sali Tagliamonte. “Not always variable: probing the vernacular grammar”. In: *Language Variation and Change* 27.3 (2015), pp. 255–285.

[78] Mark Davies. *Corpus del Español News on the Web*. 2020. (Visited on 05/01/2020).

[79] Mark Davies. “Making Google Books n-grams useful for a wide range of research on language change”. In: *International Journal of Corpus Linguistics* 19.3 (2014), pp. 401–416.

[80] David DeCamp. “Hypercorrection and rule generalization”. In: *Language in Society* 1.1 (1972), pp. 87–90.

[81] Marco Del Tredici, Raquel Fernández, and Gemma Boleda. “Short-Term Meaning Shift: A Distributional Exploration”. In: *NAACL*. 2019, pp. 2069–2075.

[82] Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández. “You Shall Know a User by the Company It Keeps: Dynamic Representations for Social Media Users in NLP”. In: *EMNLP*. 2019, pp. 4701–4711.

[83] Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. “Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings”. In: *NAACL*. 2019, pp. 2970–3005.

[84] Bianca DiJulio, Cailey Muñana, and Mollyann Brodie. *Puerto Rico after Hurricane Maria: The Public’s Knowledge and Views of Its Impact and the Response*. Tech. rep. Kaiser Family Foundation, 2017.

[85] Stewart I Donaldson and Elisa J Grant-Vallone. “Understanding self-report bias in organizational behavior research”. In: *Journal of business and Psychology* 17.2 (2002), pp. 245–260.

[86] Gabriel Doyle and Michael C Frank. “Shared common ground influences information density in microblog texts”. In: *NAACL*. 2015.

[87] Patrick Drouin and Pascaline Dury. “When terms disappear from a specialized lexicon: A semi-automatic investigation into necrology”. In: *Actes de la conférence internationale “Language for Special Purposes”*. 2009.

[88] Natasha Duarte, Emma Llanso, and Anna Loup. “Mixed Messages? The Limits of Automated Social Media Content Analysis”. In: *Conference on Fairness, Accountability and Transparency*. 2018, pp. 106–106.

[89] Line Dubé, Anne Bourhis, and Réal Jacob. “Towards a typology of virtual communities of practice”. In: *Interdisciplinary Journal of Information, Knowledge, and Management* 1.1 (2006), pp. 69–93.

[90] Antoine Dubois, Emilio Zagheni, Kiran Garimella, and Ingmar Weber. “Studying migrant assimilation through Facebook interests”. In: *SocInfo*. 2018, pp. 51–60.

[91] BK Dumas and Jonathan Lighter. “Is slang a word for linguists?” In: *American Speech* 53.1 (1978), pp. 5–17.

[92] Maeve Eberhardt and Kara Freeman. “‘First things first, I’m the realest’: Linguistic appropriation, white privilege, and the hip-hop persona of Iggy Azalea”. In: *Journal of Sociolinguistics* 19.3 (2015), pp. 303–327.

[93] Penelope Eckert. “The whole woman: Sex and gender differences in variation”. In: *Language variation and change* 1.03 (1989), pp. 245–267.

[94] Penelope Eckert. “Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation”. In: *Annual Review of Anthropology* 41 (2012), pp. 87–100.

[95] Penelope Eckert. “Variation and the indexical field”. In: *Journal of Sociolinguistics* 124 (2008), pp. 453–476.

[96] Penelope Eckert. “Variation, meaning and social change”. In: *Sociolinguistics: theoretical debates* (2016), pp. 68–85.

[97] Penelope Eckert and Sally McConnell-Ginet. “Think Practically and Look Locally: Language and Gender as Community-Based Practice”. In: *Annual Review of Anthropology* 21 (1992), pp. 461–490.

[98] Penelope Eckert and John R Rickford. *Style and sociolinguistic variation*. Cambridge University Press, 2001.

[99] Leo Egghe. “Untangling Herdan’s law and Heaps’ Law: Mathematical and informetric arguments”. In: *Journal of the American Society for Information Science and Technology* 58.5 (2007), pp. 702–709.

[100] Suzanne Eggins. *Introduction to systemic functional linguistics*. A&C Black, 2004.

[101] Jacob Eisenstein. “Identifying Regional Dialects in On-Line Social Media”. In: *The handbook of dialectology* 2013 (2013), pp. 368–383.

[102] Jacob Eisenstein. *Introduction to Natural Language Processing*. MIT Press, 2019.

[103] Jacob Eisenstein. “What to do about bad language on the internet”. In: *NAACL*. 2013, pp. 359–369.

[104] Jacob Eisenstein, Brendan O’Connor, Noah A Smith, and Eric P Xing. “A latent variable model for geographic lexical variation”. In: *EMNLP*. ACL. 2010, pp. 1277–1287.

[105] Jacob Eisenstein, Brendan O’Connor, Noah A. Smith, and Eric P. Xing. “Diffusion of lexical change in social media”. In: *PLoS ONE* 9.11 (2014).

[106] Zsuzsanna Fagyal, Samarth Swarup, Anna María Escobar, Les Gasser, and Kiran Lakkaraju. “Centers and peripheries: Network roles in language change”. In: *Lingua* 120.8 (2010), pp. 2061–2079.

[107] Kathleen Ferrara and Barbara Bell. “Sociolinguistic variation and discourse function of constructed dialogue introducers: The case of be+like”. In: *American Speech* 70.3 (1995), pp. 265–290.

[108] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. “Incorporating non-local information into information extraction systems by Gibbs sampling”. In: *ACL*. 2005, pp. 363–370.

[109] Jenny Rose Finkel and Christopher Manning. “Nested named entity recognition”. In: *EMNLP*. 2009, pp. 141–150.

[110] Terry Flew, Christina Spurgeon, Anna Daniel, and Adam Swift. “The promise of computational journalism”. In: *Journalism Practice* 6.2 (2012), pp. 157–171.

[111] Alasdair Fotheringham. “Catalan independence referendum: Region votes overwhelmingly for secession from Spain”. In: *Independent* (2017). Accessed on 30 Oct 2017.

[112] Ana Frankenberg-Garcia, Robert Lew, Jonathan C Roberts, Geraint Paul Rees, and Nirwan Sharma. “Developing a writing assistant to help EAP writers with collocations in real time”. In: *ReCALL* 31.1 (2019), pp. 23–39.

[113] Maximiliane Frobenius. “Audience design in monologues: How vloggers involve their viewers”. In: *Journal of Pragmatics* 72 (2014), pp. 59–72.

[114] Susan R Fussell and Robert M Krauss. “Understanding friends and strangers: The effects of audience design on message comprehension”. In: *European Journal of Social Psychology* 19.6 (1989), pp. 509–525.

[115] Devin Gaffney and J. Nathan Matias. “Caveat emptor, computational social science: Large-scale missing data in a widely-published reddit corpus”. In: *PLoS ONE* 13.7 (2018).

[116] Alexia Galati and Susan Brennan. “Attenuating information in spoken communication: For the speaker, or for the addressee?” In: *Journal of Memory and Language* 62.1 (2010), pp. 35–51.

[117] Huiji Gao, Jalal Mahmud, Jilin Chen, Jeffrey Nichols, and Michelle Zhou. “Modeling user attitude toward controversial topics in online social media”. In: *ICWSM*. 2014.

[118] William Gardner, Edward P Mulvey, and Esther C Shaw. “Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models”. In: *Psychological bulletin* 118.3 (1995), pp. 392–404.

[119] Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. “Women’s syntactic resilience and men’s grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing”. In: *ACL*. 2019, pp. 3493–3498.

[120] Aparna Garimella, Carmen Banea, and Rada Mihalcea. “Demographic-aware word associations”. In: *EMNLP*. 2017, pp. 2285–2295.

[121] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. “The Effect of Collective Attention on Controversial Debates on Social Media”. In: *WebSci*. 2017.

[122] Matt Garley and Julia Hockenmaier. “Beefmoves: dissemination, diversity, and dynamics of English borrowings in a German hip hop forum”. In: *ACL*. 2012, pp. 135–139.

[123] Dirk Geeraerts. “Cultural models of linguistic standardization”. In: *Cognitive models in language and thought. Ideology, metaphors and meanings*. Ed. by René Dirven, Roslyn Frank, and Martin Pütz. Vol. 2568. 2003.

[124] Clifford Geertz. “Deep play: Notes on the Balinese cockfight”. In: *Culture and Politics*. Springer, 2000, pp. 175–201.

[125] Eric Gilbert. “Widespread Underprovision on Reddit”. In: *CSCW*. 2013, pp. 803–808.

[126] Howard Giles, Nikolas Coupland, and Justine Coupland. “Accommodation theory: Communication, context, and consequence”. In: *Contexts of accommodation: Developments in applied sociolinguistics*. 1991.

[127] Kevin Gimpel, Nathan Schneider, Brendan O’Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah Smith. “Part-of-speech tagging for Twitter: Annotation, features, and experiments”. In: *ACL*. 2011, pp. 42–47.

[128] Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. “Analysing Lexical Semantic Change with Contextualised Word Representations”. In: *ACL*. 2020, pp. 3960–3973.

[129] Benjamin Gleason. “#OccupyWallStreet: Exploring informal learning about a social movement on Twitter”. In: *American Behavioral Scientist* 57.7 (2013), pp. 966–982.

[130] Kristina Gligorić, Ashton Anderson, and Robert West. “How constraints affect content: The case of Twitter’s switch from 140 to 280 characters”. In: *ICWSM*. 2018.

[131] Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. “The social dynamics of language change in online networks”. In: *SocInfo*. Springer. 2016, pp. 41–57.

[132] Erving Goffman. *The presentation of self in everyday life*. Harmondsworth London, 1978.

[133] Felix Rodriguez Gonzalez. “Anglicisms in contemporary Spanish. An overview”. In: *Atlantis* 21.1/2 (1999), pp. 103–139.

[134] Kay González-Vilbazo and Luis López. “Some properties of light verbs in code-switching”. In: *Lingua* 121.5 (2011), pp. 832–850.

[135] Government of Catalonia. *Language Use of the Population of Catalonia*. Tech. rep. Accessed on 30 Oct 2017. 2013.

[136] Mark S Granovetter. “The strength of weak ties”. In: *American journal of sociology* 78.6 (1973), pp. 1360–1380.

[137] Lisa J Green. *African American English: a linguistic introduction*. Cambridge University Press, 2002.

[138] Herbert P Grice. “Logic and conversation”. In: *Speech acts*. Brill, 1975, pp. 41–58.

[139] Jack Grieve. “Natural selection in the modern English Lexicon”. In: *International Conference on Language Evolution*. 2018, pp. 153–157.

[140] Jack Grieve, Andrea Nini, and Diansheng Guo. “Analyzing lexical emergence in Modern American English online”. In: *English Language and Linguistics* 20.2 (2016), pp. 1–29.

[141] Jack Grieve, Andrea Nini, and Diansheng Guo. “Mapping lexical innovation on American social media”. In: *Journal of English Linguistics* 46.4 (2018), pp. 293–319.

[142] John J Gumperz. “The sociolinguistic significance of conversational code-switching”. In: *RELC Journal* 8.2 (1977), pp. 1–34.

[143] John J Gumperz. “The speech community”. In: *Linguistic anthropology: A reader*. Ed. by Alessandro Duranti. 2009, pp. 381–386.

[144] Lauren Hall-Lew, Elizabeth Coppock, and Rebecca L Starr. “Indexing political persuasion: Variation in the Iraq vowels”. In: *American Speech* 85.1 (2010), pp. 91–102.

[145] William L Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. “Loyalty in online communities”. In: *ICWSM*. 2017.

[146] Eszter Hargittai. “Potential biases in big data: Omitted voices on social media”. In: *Social Science Computer Review* 38.1 (2020), pp. 10–24.

[147] Heba Hasan. “Instagram Bans Thinspo Content”. In: *Time Newsfeed* (2012).

[148] Martin Haspelmath. “Lexical borrowing: Concepts and issues”. In: *Loanwords in the world’s language: A Comparative Handbook*. 2009, pp. 944–967.

[149] Wilbert Heeringa, Peter Kleiweg, Charlotte Gooskens, and John Nerbonne. “Evaluation of string distance algorithms for dialectology”. In: *ACL*. 2006, pp. 51–62.

[150] Deborah Pacini Hernandez. *Oye como va!: Hybridity and identity in Latino Popular Music*. Temple University Press, 2010.

[151] Susan C Herring. “Grammar and electronic communication”. In: *The Encyclopedia of Applied Linguistics* (2012), pp. 1–9.

[152] Jack Hessel, Chenhao Tan, and Lillian Lee. “Science, AskScience, and BadScience: On the coexistence of highly related communities”. In: *ICWSM*. 2016, pp. 171–180.

[153] Thomas Heverin and Lisl Zach. “Use of microblogging for collective sense-making during violent crises: A study of three campus shootings”. In: *Journal of the American Society for Information Science and Technology* 63.1 (2012), pp. 34–47.

[154] Keisuke Hirano and Guido W Imbens. “The propensity score with continuous treatments”. In: *Applied Bayesian modeling and causal inference from incomplete-data perspectives*. Ed. by Andrew Gelman and Xiao-Li Meng. Chichester: Wiley, 2004, pp. 73–84.

[155] Chaya Hiruncharoenvate, Zhiyuan Lin, and Eric Gilbert. “Algorithmically Bypassing Censorship on Sina Weibo with Nondeterministic Homophone Substitutions”. In: *ICWSM*. 2015, pp. 150–158.

[156] Jake M Hofman, Amit Sharma, and Duncan J Watts. “Prediction and explanation in social systems”. In: *Science* 355.6324 (2017), pp. 486–488.

[157] Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. “Predicting the growth of morphological families from social and linguistic factors”. In: ACL. 2020.

[158] Janet Holmes and Miriam Meyerhoff. “The community of practice: Theories and methodologies in language and gender research”. In: *Language in society* 28.2 (1999), pp. 173–183.

[159] Matthew Honnibal and Mark Johnson. “An improved non-monotonic transition system for dependency parsing”. In: *EMNLP*. 2015, pp. 1373–1378.

[160] Nabil Hossain, Minh Tran, and Henry Kautz. “A Framework for Political Portmanteau Decomposition”. In: *ICWSM*. Vol. 14. 2020, pp. 944–948.

[161] J. Brian Houston, Joshua Hawthorne, Mildred F Perreault, Eun Hae Park, Marlo Goldstein Hode, Michael R Halliwell, Sarah E. Turner Mcgowen, Rachel Davis, Shivani Vaid, Jonathan A. Mcelderry, and Stanford A Griffith. “Social media and disasters: A functional framework for social media use in disaster planning, response, and research”. In: *Disasters* 39.1 (2015), pp. 1–22.

[162] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. “OntoNotes: the 90% solution”. In: *NAACL*. 2006, pp. 57–60.

[163] Xiaolei Huang and Michael Paul. “Examining temporality in document classification”. In: *ACL*. 2018, pp. 694–699.

[164] Guido W Imbens. “The role of the propensity score in estimating dose-response functions”. In: *Biometrika* 87.3 (2000), pp. 706–710.

[165] Rika Ito and Sali Tagliamonte. “Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers”. In: *Language in Society* 32 (2003), pp. 257–279.

[166] Alexandra Jaffe. “Codeswitching and stance: Issues in interpretation”. In: *Journal of Language, Identity, and Education* 6.1 (2007), pp. 53–77.

[167] Alexandra Jaffe, Jannis Androutsopoulos, Mark Sebba, and Sally Johnson. *Orthography as Social Action: Scripts, Spelling, Identity and Power*. Berlin: Walter de Gruyter, 2012.

[168] Christine Jeavens. “In maps: How close was the Scottish referendum vote?” In: *BBC News* (2014). Accessed on 30 Oct 2017.

[169] Devin L. Jenkins. “Bilingual Verb Constructions in Southwestern Spanish”. In: *Bilingual Review* (2003), pp. 195–204.

[170] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R Brubaker, and Casey Fiesler. “Moderation Challenges in Voice-based Online Communities on Discord”. In: *Proceedings of the ACM on Human-Computer Interaction* 3 (2019), pp. 1–23.

[171] Anders Johannsen, Dirk Hovy, and Anders Søgaard. “Cross-lingual syntactic variation over age and gender”. In: *CoNLL*. 2015, pp. 103–112.

[172] Benjamin K Johnson and Giulia Ranzini. “Click here to look clever: Self-presentation via selective sharing of music and film on social media”. In: *Computers in Human Behavior* 82 (2018), pp. 148–158.

[173] Barbara Johnstone, Jennifer Andrus, and Andrew E Danielson. “Mobility, indexicality, and the enregisterment of “Pittsburghese””. In: *Journal of English linguistics* 34.2 (2006), pp. 77–104.

[174] Patrick Juola. “Using the Google N-Gram corpus to measure cultural complexity”. In: *Literary and linguistic computing* 28.4 (2013), pp. 668–675.

[175] Yoonjung Kang. “Loanword phonology”. In: *The Blackwell companion to phonology* (2011), pp. 1–25.

[176] Ankit Kariryaa, Isaac Johnson, Johannes Schöning, and Brent Hecht. “Defining and Predicting the Localness of Volunteered Geographic Information using Ground Truth Data”. In: *CHI*. 2018.

[177] Andres Karjus, Richard A Blythe, Simon Kirby, and Kenny Smith. “Topical advection as a baseline model for corpus-based lexical dynamics”. In: *SCiL* 1.1 (2018), pp. 186–188.

[178] Martha Sif Karrebæk. “Don’t speak like that to her!: Linguistic minority children’s socialization into an ideology of monolingualism”. In: *Journal of Sociolinguistics* 17.3 (2013), pp. 355–375.

[179] Josh Katz. *Speaking American: How Y’all, Youse, and You Guys Talk: A Visual Guide*. Houghton Mifflin Harcourt, 2016.

[180] Josh Katz. “What Music Do Americans Love the Most? 50 Detailed Fan Maps”. In: *The New York Times* (2017).

[181] Harmanpreet Kaur, Cliff Lampe, and Walter S Lasecki. “Using affordances to improve AI support of social media posting decisions”. In: *IUI*. 2020, pp. 556–567.

[182] Daniel Kershaw, Matthew Rowe, and Patrick Stacey. “Towards Modelling Language Innovation Acceptance in Online Social Networks”. In: *Proceedings of the ACM International Conference on Web Search and Data Mining* (2016), pp. 553–562.

[183] Paul Kerswill and Ann Williams. “Creating a new town koine: Children and language change in Milton Keynes”. In: *Language in society* 29.1 (2000), pp. 65–115.

[184] Amy Jo Kim. *Community building on the web: Secret strategies for successful online communities*. Peachpit Press, 2006.

[185] Suin Kim, Ingmar Weber, Li Wei, and Alice Oh. “Sociolinguistic Analysis of Twitter in Multilingual Societies”. In: *Hypertext and Social Media*. 2014, pp. 243–248.

[186] John Klein and Melvin Moeschberger. *Survival analysis: techniques for censored and truncated data*. New York: Springer Science & Business Media, 2005.

[187] Marina Kogan, Leysia Palen, and Kenneth M Anderson. “Think Local, Retweet Global: Retweeting by the Geographically-Vulnerable during Hucrricane Sandy”. In: *CSCW*. 2015, pp. 981–993.

[188] Farshad Kooti, Winter A Mason, Krishna P Gummadi, and Meeyoung Cha. "Predicting emerging social conventions in online social networks". In: *CIKM*. 2012, pp. 445–454.

[189] Farshad Kooti, Haeryun Yang, Meeyoung Cha, P Krishna Gummadi, and Winter A Mason. "The Emergence of Conventions in Online Social Networks." In: *ICWSM*. 2012.

[190] Bernd Kortmann and Benedikt Szmrecsanyi. "Parameters of morphosyntactic variation in World Englishes: prospects and limitations of searching for universals". In: *Linguistic universals and language variation* 1 (2011), pp. 264–290.

[191] Robert E Kraut and Paul Resnick. *Building successful online communities: Evidence-based social design*. Mit Press, 2012.

[192] Anthony S Kroch. "Reflexes of grammar in patterns of language change". In: *Language Variation and Change* 1.3 (1989), pp. 199–244.

[193] William Kruskal. "Relative importance by averaging over orderings". In: *The American Statistician* 41.1 (1987), pp. 6–10.

[194] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. "Statistically significant detection of linguistic change". In: *WWW*. 2015, pp. 625–635.

[195] Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. "Freshman or fresher? quantifying the geographic variation of language in online social media". In: *ICWSM*. 2016.

[196] Vivek Kulkarni and William Yang Wang. "Simple Models for Word Formation in English Slang". In: *InNAACL*. 2018, pp. 1424–1434.

[197] William Labov. *Principles of Linguistic Change: Internal Factors*. Vol. 1. Blackwell Publishers, 1994.

[198] William Labov. *Principles of Linguistic Change: Social Factors*. Vol. 2. Wiley-Blackwell, 2001.

[199] William Labov. "Some Principles of Linguistic Methodology". In: *Language in Society* 1 (1972), pp. 97–120.

[200] William Labov. "The intersection of sex and social class in the course of linguistic change". In: *Language variation and change* 2.2 (1990), pp. 205–254.

[201] William Labov. “The Social Motivation of a Sound Change”. In: *Word* 19.3 (1963), pp. 273–309.

[202] William Labov. “The social setting of linguistic change”. In: *Diachronic, Areal, and Typological Linguistics* (1973), pp. 195–249.

[203] William Labov. *The social stratification of English in New York city*. Cambridge University Press, 2006.

[204] William Labov. “Transmission and Diffusion”. In: *Language* 83.2 (2007), pp. 344–387.

[205] William Labov, Sharon Ash, and Charles Boberg. *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter, 2008.

[206] Hans J Ladegaard. “Audience design revisited: Persons, roles and power relations in speech interactions.” In: *Language & Communication* (1995).

[207] Hans J Ladegaard. “Language attitudes and sociolinguistic behaviour: Exploring attitude-behaviour relations in language”. In: *Journal of sociolinguistics* 4.2 (2000), pp. 214–233.

[208] Jean Lave and Étienne Wenger. *Situated learning: Legitimate peripheral participation*. Cambridge university press, 1991.

[209] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, and Myron Gutmann. “Life in the network: the coming age of computational social science”. In: *Science* 323.5915 (2009), p. 721.

[210] Alex Leavitt and Joshua A Clark. “Upvoting hurricane Sandy: event-based news production processes on a social news site”. In: *CHI*. 2014, pp. 1495–1504.

[211] Janette Lehmann, Bruno Gonçalves, José J Ramasco, and Ciro Cattuto. “Dynamical Classes of Collective Attention in Twitter”. In: *WWW*. 2012.

[212] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. “Meme-tracking and the dynamics of the news cycle”. In: *KDD*. ACM. 2009, pp. 497–506.

[213] Shiri Lev-Ari and Sharon Peperkamp. “An experimental study of the role of social factors in language change: The case of loanword adaptations”. In: *Laboratory Phonology* 5.3 (2014), pp. 379–401.

[214] Shiri Lev-Ari, Marcela San Giacomo, and Sharon Peperkamp. “The effect of domain prestige and interlocutors’ bilingualism on loanword adaptations”. In: *Journal of Sociolinguistics* 18.5 (2014), pp. 658–684.

[215] Vladimir Levenshtein. “Binary codes capable of correcting deletions, insertions and reversals”. In: *Soviet Physics Doklady*. Vol. 10. 1966, pp. 707–710.

[216] Y-R Lin, B Keegan, D Margolin, and D Lazer. “Rising Tides or Rising Stars? Dynamics of Shared Attention on Twitter during Media Events”. In: *PLoS ONE* 9.5 (2014).

[217] Kimberly Ling, Gerard Beenen, Pamela Ludford, Xiaoqing Wang, Klarissa Chang, Xin Li, Dan Cosley, Dan Frankowski, Loren Terveen, and Al Mamunur Rashid. “Using social psychology to motivate contributions to online communities”. In: *Journal of Computer-Mediated Communication* 10.4 (2005), pp. 00–00.

[218] John Lipski. *Latin American Spanish*. New York: Longman, 1994.

[219] John M Lipski. “Code-switching or borrowing? No sé so no puedo decir, you know”. In: *Selected proceedings of the second workshop on Spanish sociolinguistics*. Cascadilla Proceedings Project Somerville, MA. 2005, pp. 1–15.

[220] Eden Litt. “Knock, knock. Who’s there? The imagined audience”. In: *Journal of broadcasting & electronic media* 56.3 (2012), pp. 330–345.

[221] Daniel Long and Dennis R Preston. *Handbook of perceptual dialectology*. Vol. 2. John Benjamins Publishing, 2002.

[222] Bronwen Low, Mela Sarkar, and Lise Winer. ““Ch’us mon propre Bescherelle”: Challenges from the Hip-Hop nation to the Quebec nation”. In: *Journal of Sociolinguistics* 13.1 (2009), pp. 59–82.

[223] Marco Lui and Timothy Baldwin. “langid.py: An off-the-shelf language identification tool”. In: *ACL*. 2012, pp. 25–30.

[224] Yiwei Luo, Dan Jurafsky, and Beth Levin. “From insanely jealous to insanely delicious: Computational models for the semantic bleaching of English intensifiers”. In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. 2019, pp. 1–13.

[225] Brian MacWhinney. “Competition and Lexical Categorization”. In: *Linguistic Categorization* (1989), pp. 195–242.

[226] Carmen Maíz-Arévalo and Antonio García-Gómez. ““You look terrific!” Social evaluation and relationships in online compliments”. In: *Discourse Studies* 15.6 (2013), pp. 735–760.

[227] Alice E Marwick and danah boyd. “I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience”. In: *New media & society* 13.1 (2011), pp. 114–133.

[228] Corrine McCarthy. “The Northern Cities Shift in Chicago”. In: *Journal of English Linguistics* 39.2 (2011), pp. 166–187.

[229] Gretchen McCulloch. *Because Internet: Understanding the new rules of language*. Riverhead Books, 2019.

[230] Allan Metcalf. *Predicting new words: The secrets of their success*. New York: Houghton Mifflin, 2004.

[231] Jacob Metcalf and Kate Crawford. “Where are human subjects in big data research? The emerging ethics divide”. In: *Big Data & Society* 3.1 (2016), p. 2053951716650211.

[232] Katina Michael. “Bots Trending Now: Disinformation and Calculated Manipulation of the Masses”. In: *IEEE Technology and Society Magazine* 36.2 (2017), pp. 6–11.

[233] James Milroy and Lesley Milroy. “Linguistic change, social network and speaker innovation”. In: *Journal of Linguistics* 21.2 (1985), pp. 339–384.

[234] Tanushree Mitra, Graham Wright, and Eric Gilbert. “Credibility and Dynamics of Collective Attention”. In: *Proceedings of ACM on Human-Computer Interaction* 1 (2016).

[235] Luis Moreno, Ana Arriba, and Araceli Serrano. “Multiple identities in decentralized Spain: The case of Catalonia”. In: *Regional & Federal Studies* 8.3 (1998), pp. 65–88.

[236] Anthony Mulac, Karen T Erlandson, W Jeffrey Farrar, Jennifer S Hallett, Jennifer L Molloy, and Margaret E Prescott. ““Uh-huh. What’s that all about?” Differing interpretations of conversational backchannels and questions as sources of miscommunication across gender boundaries”. In: *Communication Research* 25.6 (1998), pp. 641–668.

[237] Dhiraj Murthy and Alexander J. Gross. “Social media processes in disasters: Implications of emergent technology use”. In: *Social Science Research* 63 (2017), pp. 356–370.

[238] Carol Myers-Scotton. “A theoretical introduction to the markedness model”. In: *Codes and consequences: Choosing linguistic varieties*. 1998.

[239] Innocent Ndubuisi-Obi, Sayan Ghosh, and David Jurgens. “Wetin dey with these comments? Modeling Sociolinguistic Factors Affecting Code-switching Behavior in Nigerian Online Discussions”. In: *ACL*. 2019, pp. 6204–6214.

[240] Terttu Nevalainen. “Negative Concord as an English “Vernacular Universal” Social History and Linguistic Typology”. In: *Journal of English Linguistics* 34.3 (2006), pp. 257–278.

[241] Dong Nguyen and Leonie Cornips. “Automatic detection of intra-word code-switching”. In: *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. 2016, pp. 82–86.

[242] Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. “Computational sociolinguistics: A survey”. In: *Computational Linguistics* 42.3 (2016), pp. 537–593.

[243] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. ““How old do you think I am?” A study of language and age in Twitter”. In: *ICWSM*. 2013.

[244] Dong Nguyen and Carolyn Rose. “Language use as a reflection of socialization in online communities”. In: *Workshop on Language in Social Media (LSM 2011)*. 2011, pp. 76–85.

[245] Dong Nguyen, Dolf Trieschnigg, and Leonie Cornips. “Audience and the Use of Minority Languages on Twitter”. In: *ICWSM*. 2015, pp. 666–669.

[246] James M Olson and Mark P Zanna. “Attitudes and attitude change”. In: *Annual review of psychology* 44.1 (1993), pp. 117–154.

[247] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. “What to Expect When the Unexpected Happens: Social Media Communications Across Crises”. In: *CSCW*. 2015, pp. 994–1009.

[248] Darren Paffey. “Policing the Spanish language debate: verbal hygiene and the Spanish language academy (Real Academia Española)”. In: *Language Policy* 6.3-4 (2007), pp. 313–332.

[249] Evangelos Papalexakis, Dong Nguyen, and A Seza Doğruöz. “Predicting code-switching in multilingual communication for immigrant communities”. In: *Proceedings of the first workshop on computational approaches to code switching*. 2014, pp. 42–50.

[250] Jennifer S Pardo. “On phonetic convergence during conversational interaction”. In: *The Journal of the Acoustical Society of America* 119.4 (2006), pp. 2382–2393.

[251] Jennifer S Pardo, Rachel Gibbons, Alexandra Suppes, and Robert M Krauss. “Phonetic convergence in college roommates”. In: *Journal of Phonetics* 40.1 (2012), pp. 190–197.

[252] Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. “Automatic personality assessment through social media language.” In: *Journal of personality and social psychology* 108.6 (2015), p. 934.

[253] Alan Partington. “Corpus evidence of language change: The case of the intensifier”. In: *Text and Technology: In Honour of John Sinclair*. Ed. by Mona Baker, Gill Francis, and Elena Tognini-Bonelli. Philadelphia: John Benjamins Publishing, 1993, pp. 177–192.

[254] Desmond U Patton, William R Frey, Kyle A McGregor, Fei-Tzin Lee, Kathleen McKeown, and Emanuel Moss. “Contextual Analysis of Social Media: The Promise and Challenge of Eliciting Context in Social Media Posts with Natural Language Processing”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 337–342.

[255] Michael J Paul and Mark Dredze. “You are what you tweet: Analyzing twitter for public health”. In: *ICWSM*. 2011, pp. 265–272.

[256] Umashanthi Pavalanathan and Jacob Eisenstein. “Audience-Modulated Variation in Online Social Media”. In: *American Speech* 90.2 (2015), pp. 187–213.

[257] Umashanthi Pavalanathan and Jacob Eisenstein. “Confounds and Consequences in Geotagged Twitter Data”. In: *EMNLP*. 2015, pp. 2138–2148.

[258] Umashanthi Pavalanathan and Jacob Eisenstein. “More emojis, less :) The competition for paralinguistic function in microblog writing”. In: *First Monday* (2016).

[259] Umashanthi Pavalanathan, Jim Fitzpatrick, Scott F Kiesling, and Jacob Eisenstein. “A multidimensional lexicon for interpersonal stancetaking”. In: *ACL*. 2017, pp. 884–895.

[260] Umashanthi Pavalanathan, Xiaochuang Han, and Jacob Eisenstein. “Mind Your POV: Convergence of Articles and Editors Towards Wikipedia’s Neutrality Norm”. In: *Proceedings of the ACM on Human-Computer Interaction* 2 (2018), pp. 1–23.

[261] Sai Teja Peddinti, Keith W Ross, and Justin Cappos. ""On the internet, nobody knows you're a dog": a Twitter case study of anonymity in social networks". In: *Proceedings of the second ACM conference on Online social networks*. 2014, pp. 83–94.

[262] John R Perry. "Language reform in Turkey and Iran". In: *International Journal of Middle East Studies* 17.3 (1985), pp. 295–311.

[263] Theresa Perry and Lisa D Delpit. *The real Ebonics debate: Power, language, and the education of African-American children*. Beacon Press, 1998.

[264] Janet B Pierrehumbert. "The dynamic lexicon". In: *Handbook of laboratory phonology* (2012), pp. 173–183.

[265] Bernd Ploderer, Steve Howard, and Peter Thomas. "Being online, living offline: the influence of social ties over the appropriation of social network sites". In: *CSCW*. 2008, pp. 333–342.

[266] Shana Poplack. "Sometimes I'll start a sentence in Spanish y termino en español: toward a typology of code-switching". In: *Linguistics* 18.7-8 (1980), pp. 581–618.

[267] Shana Poplack and Nathalie Dion. "Myths and facts about loanword development". In: *Language Variation and Change* 24.3 (2012), pp. 279–315.

[268] Shana Poplack and David Sankoff. "Borrowing: the synchrony of integration". In: (1984).

[269] Shana Poplack, David Sankoff, and Christopher Miller. "The social correlates and linguistic processes of lexical borrowing and assimilation". In: *Linguistics* 26 (1988), pp. 47–104.

[270] Jacob Poushter, Caldwell Bishop, and Hanyu Chwe. "Social media use continues to rise in developing countries but plateaus across developed ones". In: *Pew Research Center* 22 (2018), pp. 2–19.

[271] Jenny Preece, Blair Nonnecke, and Dorine Andrews. "The top five reasons for lurking: improving community experiences for everyone". In: *Computers in human behavior* 20.2 (2004), pp. 201–223.

[272] Dennis R Preston. "Language with an attitude". In: *The handbook of language variation and change* 40 (2002), p. 66.

[273] Dennis R Preston. "The influence of regard on language variation and change". In: *Journal of Pragmatics* 52 (2013), pp. 93–104.

[274] EF Prince. “The ZPG letter: Subjects, definiteness, and information-status”. In: *Discourse Description: Diverse linguistic analyses of a fund-raising text*. Ed. by William Mann and Sandra Thompson. 1992, pp. 295–325.

[275] Virginia Pulcini, Cristiano Furiassi, and Félix Rodríguez González. “The lexical influence of English on European languages”. In: *The Anglicization of European lexis* 1 (2012).

[276] Afshin Rahimi, Yuan Li, and Trevor Cohn. “Massively Multilingual Transfer for NER”. In: *ACL*. July 2019, pp. 151–164.

[277] Yuqing Ren, Robert Kraut, Sara Kiesler, and Paul Resnick. “Encouraging commitment in online communities”. In: *Evidence-based social design: Mining the social sciences to build online communities*. MIT Press, 2011, pp. 77–125.

[278] John Rickford and Mackenzie Price. “Girlz II women: Age-grading, language change and stylistic variation”. In: *Journal of Sociolinguistics* 17.2 (2013), pp. 143–179.

[279] John R Rickford, Arnetha Ball, Renee Blake, Raina Jackson, and Nomi Martin. “Rappin on the copula coffin: Theoretical and methodological issues in the analysis of copula variation in African-American Vernacular English”. In: *Language Variation and Change* 3.1 (1991), pp. 103–132.

[280] John R Rickford and Faye McNair-Knox. “Addressee-and topic-influenced style shift: A quantitative sociolinguistic study”. In: *Sociolinguistic perspectives on register*. Oxford: Oxford University Press, 1994, pp. 235–276.

[281] Sean Rintel. “Crisis memes: The importance of templatability to Internet culture and freedom of expression”. In: *Australasian Journal of Popular Culture* 2.2 (2013), pp. 253–271.

[282] Alan Ritter, Sam Clark, and Oren Etzioni. “Named entity recognition in tweets: an experimental study”. In: *EMNLP*. 2011, pp. 1524–1534.

[283] C Rodney and C Jubilado. “Morphological Study Verb Anglicisms in Spanish Language Morphological Study of Verb of Anglicisms in Spanish Computer Computer Language”. In: *Polyglossia* 23 (2012), pp. 43–47.

[284] Shane L Rogers, Nicolas Fay, and Murray Maybery. “Audience design through social interaction during group discussion”. In: *PLoS One* 8.2 (2013), e57211.

[285] Daniel Romero, Brendan Meeder, and Jon Kleinberg. “Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter”. In: *WWW*. ACM. 2011, pp. 695–704.

[286] Maria Ryskina, Ella Rabinovich, Taylor Berg-Kirkpatrick, David R Mortensen, and Yulia Tsvetkov. “Where New Words Are Born: Distributional Semantic Analysis of Neologisms and Their Semantic Neighborhoods”. In: *SCiL* 3.1 (2020), pp. 43–52.

[287] Matthew J Salganik. *Bit by bit: Social research in the digital age*. Princeton University Press, 2019.

[288] Kamal Salhi. “Critical imperatives of the French language in the Francophone world: Colonial legacy–postcolonial policy”. In: *Current issues in language planning* 3.3 (2002), pp. 317–345.

[289] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. “The risk of racial bias in hate speech detection”. In: *ACL*. 2019, pp. 1668–1678.

[290] Theresa Sauter and Axel Bruns. “#auspol: The hashtag as community, event, and material object for engaging with Australian politics”. In: *Hashtag publics: The power and politics of discursive networks*. Peter Lang, 2015, pp. 47–60.

[291] Dave Sayers. “The mediated innovation model: A framework for researching media influence in language change”. In: *Journal of Sociolinguistics* 18.2 (2014), pp. 185–212.

[292] Carol L Schmid. *The politics of language: Conflict, identity and cultural pluralism in comparative perspective*. New York, New York: Oxford University Press, 2001.

[293] Ana Lucía Schmidt, Fabiana Zollo, Michela Del Vicario, Alessandro Bessi, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. “Anatomy of news consumption on Facebook”. In: *Proceedings of the National Academy of Sciences* 114.12 (2017), pp. 3035–3039.

[294] Scots Language Centre. *Brief Analysis of the 2011 Census Results*. Tech. rep. Accessed on 30 Oct 2017. 2011.

[295] Hunter Shobe. “Place, identity and football: Catalonia, catalanisme and football club Barcelona, 1899-1975”. In: *National identities* 10.3 (2008), pp. 329–343.

[296] Philippa Shoemark, James Kirby, and Sharon Goldwater. “Inducing a lexicon of sociolinguistic variables from code-mixed text”. In: *Workshop on Noisy User-generated Text*. 2018, pp. 1–6.

[297] Philippa Shoemark, James Kirby, and Sharon Goldwater. “Topic and audience effects on distinctively Scottish vocabulary usage in Twitter data”. In: *Workshop on Stylistic Variation*. 2017, pp. 59–68.

[298] Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. “Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings”. In: *EMNLP*. 2019, pp. 66–76.

[299] Phillippa Shoemark, Debnil Sur, Luke Shrimpton, Iain Murray, and Sharon Goldwater. “Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media”. In: *EACL*. 2017, pp. 1239–1248.

[300] Julia Snell. “From sociolinguistic variation to socially strategic stylisation”. In: *Journal of Sociolinguistics* 14.5 (2010), pp. 630–656.

[301] Robert Soden and Leysia Palen. “Informating crisis: Expanding critical perspectives in crisis informatics”. In: *Proceedings of ACM on Human-Computer Interaction* 2 (2018), pp. 1–22.

[302] Thamar Solorio and Yang Liu. “Learning to predict code-switching points”. In: *EMNLP*. 2008, pp. 973–981.

[303] Lauren Squires. “Enregistering internet language”. In: *Language in Society* 39 (2010), pp. 457–492.

[304] Ieva Staliūnaitė, Hannah Rohde, Bonnie Webber, and Annie Louis. “Getting to “Hearer-old”: Charting Referring Expressions Across Time”. In: *EMNLP*. 2018, pp. 4350–4359.

[305] James N Stanford. “A call for more diverse sources of data: Variationist approaches in non-English contexts”. In: *Journal of Sociolinguistics* 20.4 (2016), pp. 525–541.

[306] Ian Stewart, Stevie Chancellor, Munmun De Choudhury, and Jacob Eisenstein. “#Anorexia, #anarexia, #anarexyia: Characterizing online community practices with orthographic variation”. In: *2017 IEEE International Conference on Big Data (Big Data)*. IEEE. 2017, pp. 4353–4361.

[307] Ian Stewart and Jacob Eisenstein. “Making “fetch” happen: The influence of social and linguistic context on nonstandard word growth and decline”. In: *EMNLP*. 2018, pp. 4360–4370.

[308] Ian Stewart, René D Flores, Timothy Riffe, Ingmar Weber, and Emilio Zagheni. “Rock, Rap, or Reggaeton?: Assessing Mexican Immigrants’ Cultural Assimilation Using Facebook Data”. In: *WWW*. 2019, pp. 3258–3264.

[309] Ian Stewart, Yuval Pinter, and Jacob Eisenstein. “Si O No, Que Penses? Catalonian Independence and Linguistic Identity on Social Media”. In: *NAACL*. 2018, pp. 136–141.

[310] Ian Stewart, Diyi Yang, and Jacob Eisenstein. “Characterizing Collective Attention via Descriptor Context: A Case Study of Public Discussions of Crisis Events”. In: *ICWSM*. Vol. 14. 2020, pp. 650–660.

[311] Leo Stewart, Ahmer Arif, A Conrad Nied, Emma S Spiro, and Kate Starbird. “Drawing the Lines of Contention: Networked Frame Contests Within #BlackLivesMatter Discourse”. In: *Proceedings of the ACM on Human-Computer Interaction*. 2017.

[312] Stefan Stieglitz, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. “Social media analytics–Challenges in topic discovery, data collection, and data preparation”. In: *International journal of information management* 39 (2018), pp. 156–168.

[313] Jane Stuart-Smith, Gwilym Pryce, Claire Timmins, and Barrie Gunter. “Television can also be a factor in language change: Evidence from an urban dialect”. In: *Language* (2013), pp. 501–536.

[314] Jane Stuart-Smith and Claire Timmins. “The role of the individual in language variation and change”. In: *Language and identities* (2010), pp. 39–54.

[315] Sali Tagliamonte and Derek Denis. “Linguistic ruin? LOL! Instant messaging and teen language”. In: *American Speech* 83.1 (2008), pp. 3–34.

[316] Sali A. Tagliamonte and Alexandra D’Arcy. “Frequency and variation in the community grammar: Tracking a new change through the generations”. In: *Language Variation and Change* 19.02 (2007), pp. 199–217.

[317] Sali A Tagliamonte and Jennifer Smith. “Layering, competition and a twist of fate: Deontic modality in dialects of English”. In: *Diachronica* 23.2 (2006), pp. 341–380.

[318] Chenhao Tan. “Tracing community genealogy: how new communities emerge from the old”. In: *ICWSM*. 2018.

[319] Chenhao Tan and Lillian Lee. “All who wander: On the prevalence and characteristics of multi-community engagement”. In: *WWW*. 2015, pp. 1056–1066.

[320] Rachael Tatman. ““I’ma spawts guay”: Comparing the Use of Sociophonetic Variables in Speech and Twitter”. In: *University of Pennsylvania Working Papers in Linguistics* 22.2 (2016), p. 18.

[321] Mariona Taulé, Maria Antònia Martí, and Marta Recasens. “AnCora: Multilevel Annotated Corpora for Catalan and Spanish.” In: *LREC*. 2008, pp. 96–101.

[322] Kyla Thomas. “Sounds of disadvantage: Musical taste and the origins of ethnic difference”. In: *Poetics* 60 (2017), pp. 29–47.

[323] Salah G Thomason. *Language contact*. Citeseer, 2001.

[324] Scott Tonidandel, James LeBreton, and Jeff Johnson. “Determining the statistical significance of relative weights.” In: *Psychological methods* 14.4 (2009), p. 387.

[325] Marco Del Tredici and Raquel Fernández. “The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities”. In: *COLING*. 2018, pp. 1591–1603.

[326] HC Triandis. “Attitude and attitude change”. In: *Encyclopedia of Human Biology*. Ed. by Renato Dulbecco. Vol. 1. 1991, pp. 485–496.

[327] Peter Trudgill. “Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography”. In: *Language in society* 3.2 (1974), pp. 215–246.

[328] Natsuko Tsujimura and Stuart Davis. “A construction approach to innovative verbs in Japanese”. In: *Cognitive Linguistics* 22.4 (2011), pp. 799–825.

[329] Oren Tsur and Ari Rappoport. “Don’t Let Me Be #Misunderstood: Linguistically Motivated Algorithm for Predicting the Popularity of Textual Memes”. In: *ICWSM*. 2015, pp. 426–435.

[330] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. “Social media, political polarization, and political disinformation: A review of the scientific literature”. In: *Hewlett Foundation* (2018).

[331] István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. “Aid is Out There: Looking for Help from Tweets during a Large Scale Disaster”. In: *ACL*. 2013, pp. 1619–1629.

[332] Shoko Wakamiya, Émilien Antoine, Adam Jatowt, Yukiko Kawai, and Toyokazu Akiyama. “Portraying Collective Spatial Attention in Twitter”. In: *KDD*. 2015, pp. 39–48.

[333] Dong Wang, Boleslaw K Szymanski, Tarek Abdelzaher, Heng Ji, and Lance Kaplan. “The age of social sensing”. In: *Computer* 52.1 (2019), pp. 36–45.

[334] Zijian Wang, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flöck, and David Jurgens. “Demographic inference and

representative population estimates from multilingual social media data". In: *WWW*. 2019, pp. 2056–2067.

[335] Samuel F Way, Santiago Gil, Ian Anderson, and Aaron Clauset. "Environmental Changes and the Dynamics of Musical Identity". In: *ICWSM*. Vol. 13. 2019, pp. 527–536.

[336] Uriel Weinreich, William Labov, and Marvin Herzog. *Empirical foundations for a theory of language change*. University of Texas Press, 1968.

[337] Max W Wheeler. "Catalan". In: *The Romance Languages*. Ed. by Martin Harris and Nigel Vincent. Taylor & Francis, 1997.

[338] Stefan Wojcik and Adam Hughes. "Sizing up Twitter users". In: *Washington, DC: Pew Research Center* (2019).

[339] Walt Wolfram and Erik Thomas. *The Development of African American English*. John Wiley & Sons, 2008.

[340] Ian D Wood. "Community topic usage in social networks". In: *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*. 2015, pp. 3–9.

[341] Zach Wood-Doughty, Praateek Mahajan, and Mark Dredze. "Johns Hopkins or johnny-hopkins: Classifying Individuals versus Organizations on Twitter". In: *Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*. 2018, pp. 56–61.

[342] Jiang Yang, Xiao Wei, Mark S Ackerman, and Lada A Adamic. "Activity Lifespan: An Analysis of User Survival Patterns in Online Knowledge Sharing Communities." In: *ICWSM* (2010), pp. 186–193.

[343] Yi Yang and Jacob Eisenstein. "Overcoming language variation in sentiment analysis with social attention". In: *TACL* 5 (2017), pp. 295–307.

[344] Yi Yang and Jacob Eisenstein. "Unsupervised multi-domain adaptation with feature embeddings". In: *NAACL*. 2015, pp. 672–682.

[345] H Peyton Young. "The dynamics of social innovation". In: *Proceedings of the National Academy of Sciences* 108.Supplement 4 (2011), pp. 21285–21291.

[346] Rodrigo Zamith and Seth C Lewis. "Content analysis and the algorithmic coder: What computational social science means for traditional modes of media analysis". In: *The ANNALS of the American Academy of Political and Social Science* 659.1 (2015), pp. 307–318.

[347] Raffaella Zanuttini, Jim Wood, Jason Zentz, and Laurence Horn. “The Yale Grammatical Diversity Project: Morphosyntactic variation in North American English”. In: *Linguistics Vanguard* 1 (2018).

[348] Eline Zenner, Dirk Speelman, and Dirk Geeraerts. “A sociolinguistic analysis of borrowing in weak contact situations: English loanwords and phrases in expressive utterances in a Dutch reality TV show”. In: *International Journal of Bilingualism* 19.3 (2015), pp. 333–346.

[349] Eline Zenner, Dirk Speelman, and Dirk Geeraerts. “Cognitive Sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch”. In: *Cognitive Linguistics* 23.4 (2012), pp. 749–792.

[350] Justine Zhang, William L Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. “Community identity and user engagement in a multi-community landscape”. In: *ICWSM*. 2017.

[351] Justine Zhang, James Pennebaker, Susan Dumais, and Eric Horvitz. “Configuring Audiences: A Case Study of Email Communication”. In: *Proceedings of the ACM on Human-Computer Interaction* 4 (2020), pp. 1–26.

[352] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net”. In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.