UNIVERSITY OF CALIFORNIA, MERCED

Quantifying Context and its Effects in Large Natural Datasets

A dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Cognitive and Information Sciences

by

David W. Vinson

Committee in charge:

Professor Teenie Matlock, Chair

Professor Rick Dale

Professor Jeffrey Yoshimi

2017

The dissertation of David W. Vinson is approved, and it is acceptable

in quality and form for publication on microfilm and electronically:

Professor Jeff Yoshimi

Professor Rick Dale

Professor Teenie Matlock, Chair

University of California, Merced

2017

To my family,
Who believed in me long before I knew where I was going

Contents

List of Figures

List of Tables

Acknowledgements

This dissertation was shaped by many. In the process, they shaped me. I'm grateful for the opportunity to learn from my mentors including Gordon Redding for his initial interest. J. Scott Jordan who helped me develop a sense of coherence and a metaphysics I strive to live by. Rick Dale, my closest mentor and friend, who pushed me to limits I did not know I could even imagine. Teenie Matlock for her guidance, kind words and persistent encouragement. Jeff Yoshimi for his introspective and thoughtful reading of both the work and me. The Cognitive and Information Sciences Graduate Group at the University of California, Merced for its financial support and many opportunities as well as its amazing graduate students and faculty who have become my close friends. Drew Abney, Jason (Davis) Dark and Devin Gill for the late night conversations and napkin philosophy. I would like to thank those who have been so kind as to let me mentor them: Emmanuel Villanueva, Kristina Castellanos and Mayumi Amargo.

I would like to thank my family. Thank you Kelley, Keiko, Mike and Kristine for all your support. Katie, Jim, Dan, Josh, Michelle & Brittany for indulging me on my more philosophical tangents, impromptu visits, very early morning phone calls and continued support. Mom and Dad, thank you for all you have done. Especially your patience with me, your belief in me, and your persistent encouragement and love. Finally, Chelsea, thank you for being by my side through it all. Thank you for making this adventure truly wonderful.

- - -

The text of Chapter 2 of this dissertation is a partial reprint of the material as it appears in "Valence Weakly Constrains the Information density of messages", co-authored by Rick Dale. Chapter 3 is a partial reprint of the material as it appears in "Social Structure Relates to Linguistic Information Density", which was co-authored by Rick Dale supported in part by the NSF grant INSPIRE-1344279. The text of Chapter 4 is a partial reprint of the material as it appears in "Efficient n-gram analysis with cmscu", which was co-authored by Jason K, Davis (Dark), Rick Dale and Suzanne S. Sindi. The text of The text of Chapter 5 is a partial reprint of the material as it appears in "Statistics in the Fingertips: Typing Reveals Word Predictability", which was co-authored by Stephanie S. Shih, Rick Dale and Michael Spivey. The text of Chapter 6 is a partial reprint of the material as it appears in "Decision Contamination in the Wild: Sequential Dependencies in online review ratings", which was co-authored by Rick Dale and Michael N. Jones supported in part by NSF BCS-1056744 and an IBM PhD fellowship.

Curriculum Vita

David W. Vinson

PhD Candidate, Cognitive and Information Sciences
University of California, Merced

email: dave@davevinson.com
website: davevinson.com

Education

Formal Training

2012–2017	PHD in Cognitive and Information Sciences, University of California, Merced
2010–2012	MS in Cognitive and Behavioral Sciences, Illinois State University
2006–2010	BS in Psychology & Philosophy, Illinois State University

Additional Training

May-August 2017	INTERN. International Business Machines, Research. *Alamden, CA.*
August 2014	WORKSHOP. Dynamics of Joint-Action and Social Behavior: Theory, modeling and Research. *Humbodlt Universitat, Berlin*
June 2013	SUMMER SCHOOL. DAPA Advanced Training Institute: Data Mining for Psychological Science. *Davis California.*
June 2011	SUMMER SCHOOL. APA Advanced Training Institute: Nonlinear Methods in Psychological Research. *Cincinnati, Ohio.*

Fellowships & Awards

2017	Dean's Dissertation Fellowship	$18,576.68
2015-2016	IBM PhD. Fellowship	$30,000
2014	Grand Prize, Yelp Challenge	$6,500
2014	Koostama, Summer Data Science Fellow	$2,500

Publications

Technical Packages

Davis, J. D. & **Vinson, D. W.** (*Submitted*). cmscu: A Count-Min-Sketch with Conservative Update implementation for R. *R package version 1.0*

Refereed Articles

Dale, R., Galati, A., Alviar, C., Contreras Kallens, P., Ramirez-Aristizabal, A., Tabatabaeian, M., & **Vinson, D. W.**. (*Submitted*). Perspective-taking is complex and dynamic. *Frontiers in Psychology*.

Vinson, D. W., Shih, S., & Dale, R. (*Submitted*). Lexical effects on keystroke-level dynamics reflect word predictability. *Cognition*.

Vinson, D. W., Dale, R., & Jones, M. N. (*In Revision*). Decision contamination in the wild: Sequential dependencies in online review ratings. *Psychological Science*.

Vinson, D. W., Abney, D.H., Amso, D., Anderson, M. L., Chemero, T., Cutting, J.E., Dale, R., Richardson, D., Feldman, L., Freeman, J., Friston, K., Gallagher, S., Jordan, J.S., Mudrik, L., Ondobaka, S., Shams, L., Shiffrar, M., & Spivey, M. (2016). Perception, as You Make It. Commentary on C. Firestone & B. Scholl's "Cognition does not affect perception: Evaluating the evidence for 'top-down' effects". *Behavioral and Brain Sciences*, 39

Vinson, D. W., & Dale, R. (2016). Social structure relates to linguistic information density. Chapter to appear in Jones, M. N. (Ed.). Big Data in Cognitive Science: From Methods to Insights. *Psychology Press: Taylor & Francis*.

Vinson, D. W., Davis, J. K., Sindi, S. S., & Dale, R. (2016). Efficient n-gram analysis with cmscu. *Behavior Research Methods*. doi: 10.3758/s13428-016-0766-5

Vinson, D. W., Matlock, T., Zwaan, R., & Dale, R. (2016). Implied motion language can influence visual spatial memory. *Memory and Cognition*.

Jordan, J.S., & **Vinson, D. W.** (2016). Conflicts everywhere! Perceptions, actions, and cognition all entail memory and reflect conflict. Commentary on E. Morsella's "Homing in on Consciousness in the Nervous System: An Action-Based Synthesis". *Behavioral and Brain Sciences*, 39.

Vinson, D. W., Dale, R., & Jones, M. (2016). Decision contamination in the wild: Sequential Dependencies in Yelp review ratings. In D. Grodner, D. Mirman, A. Papafragou, J. Trueswell, J. Nock, S. Arunachalam, S. Christie, & C. Norris (Eds.). *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 1433-1438). Chicago

Vinson, D. W., Jordan, J.S., & Hund, A. M. (2015). Perceptually Walking in Another's Shoes: Goals and Memories Constrain Spatial Perception. *Psychological Research*. 79(5), 1-9. doi: 10.1007/s00426-015-0714-5

Vinson, D. W., Jordan, J.S., & Hund, A. M. (2015). Spatial Perception is Continuously Constrained by Goals and Memories. In R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle & J. Yoshimi (Eds.). *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 2493-2498). Chicago

Vinson, D. W., Dale, R., Tabatabaeian, M, & Duran, N.D. (2015). Seeing and believing: Social influences on language processing. In Attention and Vision in Language Processing (pp. 197-213). *Springer India*.

Abney, D.H., McBride, D.M., Conte, A., & **Vinson, D. W.** (2015). Response dynamics in prospective memory: Velocity profiles reflect cue focality. *Psychonomic Bulletin & Review*. 22, 1020-1028. doi: 10.3758/s13423-014-0771-6

Yoshimi J., & **Vinson, D. W.** (2015). Extending Gurwitsch's Field Theory of Consciousness. *Consciousness and Cognition*. 34, 104-123.

Vinson, D. W., & Dale, R. (2014). An exploration of semantic tendencies in review ratings. *In the Proceedings of the Inaugural Science and Information Conference (SAI)*, (pp.803-809). IEEE. doi: 10.1109/SAI.2014.6918278

Vinson, D. W.,& Dale, R. (2014). Valence weakly constrains the information density of messages. In P. Bello, M. Guarini, M. McShane & B. Scassellati (Eds.). *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 1682-1687). Austin,TX: Cognitive Science Society.

Vinson, D. W., Abney, D.H., Dale, R.,& Matlock, T. (2014) High-level context effects on spatial displacement: the effects of body orientation on language and memory. *Frontiers in Psychology*. 5:637. doi:10.3389/fpsyg. 2014.00637

Dale, R.,& **Vinson, D. W.** (2013). The Observer's Observer's Paradox. *Journal Of Experimental And Theoretical Artificial Intelligence*, 25(3), 303-322.

Jordan, J.S., &**Vinson, D. W.**, (2012). After nature: On bodies, consciousness, and causality. *Journal of Consciousness Studies*, 19, 229-250.

Redding, G. M., & **Vinson, D. W.** (2010). Virtual and drawing structures for the Müller-Lyer illusions. *Attention, Perception and Psychophysics*, 72(5), 1350-1366

Published Abstracts

Bernal, R., Dale, R., & **Vinson, D. W.**, (2017, Sept) The Relationship of Social Network Connectivity to Positive Emotion Word Use and Other LIWC Word Categories. *Proceedings of the 14th annual Conference on Complex Systems*, Torino, Italy.

Vinson, D. W., (2016, Nov). Using complicated n-gram smoothing algorithms efficiently in R. *Proceedings of the 46th annual meeting of the Society for Computers in Psychology*, Boston, MA, USA.

Feldman, L.B., **Vinson, D. W.**, & Dale, R. (2016, Nov). Production of morphologically complex words as revealed by a typing task: Keystroke dynamics on the (same) stem change with whole word frequency from affixation. *Proceedings of the 57th annual meeting of the Psychonomics Society*. Boston, Massachusetts, USA.

Vinson, D. W., Dale, R., & Jones, M. (2016, June). Decision contamination in the wild: Sequential Dependencies in Yelp review ratings. *In Proceedings of the 26th Annual Meeting of the Canadian Society for Brain Behavior and Cognitive Science*. Ottawa, Canada.

Vinson, D. W., Davis, J., Sindi, S., & Dale., R. (2015, Nov). Echoes of social and cognitive constraints on language use in a large naturally occurring data set. *Proceedings of the 45th annual meeting of the Society for Computers in Psychology*, Chicago, Illinois, USA.

Davis, J., **Vinson, D. W.**, Sindi, S., & Dale., R. (2015, Nov). The transformative value of high-performance computing to big data research in cognitive science. *Proceedings of the 45th annual meeting of the Society for Computers in Psychology*, Chicago, Illinois, USA.

Vinson, D. W. & Dale., R. (2015, Nov). Message Valence Relates to Language Structure in Online Reviews. *Proceedings of the 56th annual meeting of the Psychonomics Society*. Chicago, Illinois, USA.

Vinson, D. W. & Dale, R. (2015, July). Communication in Communities: Exploring Subtle Influences on Language Use in More than a Million Online Reviews. In the J. Vandekerckhove & J. Trueblood (Eds.) *Proceedings of the 48th Annual Meeting of the Society for Mathematical Psychology*, Newport Beach, CA.

Vinson, D. W. & Dale, R. (2015, July). Linguistic structure adapts to social structure. *Proceedings of the 4th Biannual Meeting of the Society for Complex Systems in Cognitive science*, Pasadena, CA.

Vinson, D. W. & Dale, R. (2015, July). Social Network Structure Contributes to Differences in Language use. In R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle & J. Yoshimi (Eds.) *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin,TX: Cognitive Science Society.

Vinson, D. W., Zwaan, R., Dale, D., Matlock, T. & Engelen, J. (2014, Nov). Motion language mediates gravitational effects on spatial memory. *Proceedings of the 12th incarnation of the Cognitive Linguistics and Discourse Processing*. Santa Barbara, California, USA.

Vinson, D. W. & Dale, R. (2014, Nov). Information and social network analysis in a large natural corpus. *Proceedings of the 44th annual meeting of the Society for Computers in Psychology*, Long Beach, CA, USA.

Vinson, D. W., Dale, R. (2014, Nov). Message valence constrains information density. *Proceedings of the 55th annual meeting of the Psychonomics Society*. Long Beach, California, USA.

Vinson, D. W., Zwaan, R., Dale, R., Matlock, T., & Engelen, J.(2014, Aug) Visual evidence that simulating language occurs in realistic observation. *Proceedings of the 7th annual conference of Embodied and Situated Language Processing*. Rotterdam, Netherlands.

Vinson, D. W., & Dale, R. (2014, June). Message valence constrains information density. *26th American Psychological Science Annual Convention*, May 22-25, San Francisco, California, USA.

Vinson, D. W. & Abney, D.H. (2013, Aug.). The influence of linguistic and social information on visual memory. *Proceedings of the 2nd Annual Cognition and Language workshop*. Santa Barbara,California, USA.

Vinson, D.W., & Dale, R. (2013, Aug.). How CRQA can shed light on two person anticipatory systems. *Proceedings of the 5th International Recurrence Plot Symposium*. Chicago, Illinois, USA.

Vinson, D. W. & Dale, R. (2013, July). Social gaze orientation influences decision making. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society. Berlin, Germany.

Vinson, D. W., Abney, D.H., Dale, R. & Matlock, T. (2013, July). The influence of motion language and gaze orientation on spatial memory. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society. Berlin, Germany.

Vinson, D. W., Abney, D. H., Dale, R. & Matlock, T. (2013, July). Visuospatial memory's sensitivity to language and image orientation. *Proceedings of the 6th annual conference of Embodied and Situated Language Processing*. Potsdam, Germany.

Vinson, D. W. & Jordan, J.S. (2013, July). Who carries your past? How social contexts and remembered actions influence perceived distance. *Proceedings of the 5th biannual Joint Action Meeting*. Berlin, Germany.

Vinson, D. W. & Jordan, J.S. (2012, Nov). Perceived distance is influenced by memory and social factors. *Proceedings of the 53rd annual meeting of the Psychonomic Society* (p. 65). Minneapolis, Minnesota, USA.

Vinson, D. W. & Jordan, J.S. (2011, Nov.). Social Proffitt: the influence of burden dynamics of another. *Proceedings of the 52nd annual meeting of the Psychonomics Society* (p140). Seattle, Washington, USA.

Vinson, D. W. & Jordan, J.S. (2011, Sept.). The Social influence of action-effect contingencies. *Illinoisy Data Conference*. Carbondale, Illinois, USA.

Vinson, D. W. & Jordan, J.S. (2010, April). Social perception of action constraints. *ILLOWA: Undergraduate Conference*. Galesberg, Illinois, USA.

Media

(2016, Oct). Sequential Dependence in online restaurant reviews. *www.dataonthemind.org/node/1600*

Jennings, C.D., Cobb, P., Vinson, D. W. (2016, May, 3). Academic Placement Data and Analysis: An Update with a Focus on Gender. *http://blog.apaonline.org/2016/05/03/academic-placement-data-and-analysis-anupdate-with-a-focus-on-gender/*

Snyder, C. (2015, Mar, 25) Cognitive science student earns IBM Ph.D. Fellowship. University of California, Merced, University News.

(2014, Mar, 11). UC Merced Connect: Study shows link between attitude and word choice. Merced Sun-Star.

Hernandez-Jason, S. (2014, Mar, 3). Research shows connections between attitudes, language. University of California, Merced, University News

Presentations

Invited Oral Presentations

November 2016	International Business Machines, Almaden Research Center	San Jose, CA.
April 2014	Language and Cognition lab, Department of Psychology, Stanford University	Palo Alto, CA.
March 2015	CogNetwork, Department of Linguistics, University of California, Berkeley	Berkeley, CA.
January 2015	Cognitive and Behavioral Science Pro Seminar, Illinois State University	Normal IL.
November 2014	IBM Almaden Research Center	San Jose, CA.
August 2014	Yelp, Inc.	San Francisco, CA.

Oral Presentations

November 2016	57th Annual meeting of the Psychonomics Society	Chicago, IL.
November 2016	46h Annual meeting of the Society for Computers in Psychology	Boston, MA.

November 2015	45th Annual meeting of the Society for Computers in Psychology	Long Beach, CA.
July 2015	48th Annual Meeting of the Society for Mathematical Psychology	Newport Beach, CA.
July 2015	4th Biannual meeting of the Society for Complex Systems in Cognitive Science	Pasadena CA.
November 2015	12th Incarnation of the Cognitive Linguistics and Discourse Processing	Santa Barbara CA.
November 2015	44th Annual meeting of the Society for Computers in Psychology	Long Beach, CA.
August 2014	2nd Annual Science and Information Conference	London, UK.
August 2014	7th Annual conference of Embodied and Situated Language Processing	Rotterdam, Netherlands
July 2014	36th Annual Meeting of the Cognitive Science Society	Quebec City, Canada
August 2013	2nd Annual Cognition and Language workshop	Santa Barbara, CA.
August 2013	5th International Recurrence Plot Symposium	Chicago, IL.
September 2013	Illinoisy Data Conference	Carbondale, IL.
April 2010	ILLOWA	Galesberg, IL.

Poster Presentations

July 2015	38th Annual Meeting of the Cognitive Science Society	Quebec City, Canada
August 2016	38th Annual Meeting of the Cognitive Science Society	Philadelphia, PA.
November 2015	56th Annual meeting of the Psychonomics Society	Chicago, IL.
July 2015	37th Annual Meeting of the Cognitive Science Society	Pasadena, CA.
November 2014	55th Annual meeting of the Psychonomics Society	Long Beach, CA.
June 2014	26th APS Annual Convention	San Francisco, CA.
July 2013	2nd Annual Science and Information Conference	London, UK.
July 2013	35th Annual Meeting of the Cognitive Science Society	Berlin, Germany
July 2013	36th Annual Meeting of the Cognitive Science Society	Quebec City, Canada
July 2013	5th Biannual Joint Action Meeting	Berlin, Germany
July 2013	6th Annual conference of Embodied and Situated Language Processing	Potsdam, Germany
November 2012	53rd Annual meeting of the Psychonomic Society	Minneapolis, MN.
November 2011	52nd Annual meeting of the Psychonomics Society	Seattle, WA.

Teaching

Assisted in teaching 9 different classes across 2 different universities and 3 departments, including labs & seminars with groups of 30+, Evaluated writing, projects and examinations for up to 180 students at a time. Topics include *Cognitive Science, Philosophy, Psychology, Linguistics, Statistics, Research methods, Logic*.

Service

Editor

2017–Present	Behavioral Research Methods

Workshops

June-2017	Data on the Mind Workshop Instructor, University of California, Berkeley	Berkeley Ca.
February-2017	Dynamics of Language Instructor, University of California, Santa Barbara	Santa Barbara, Ca.

Ad Hoc Reviewer

2015–Present	PLoS ONE
2014–Present	Behavior Research Methods
2012–Present	The Annual Meeting of the Cognitive Science Society
2014–2015	Mishra, R. K., Srinivasan, N., & Huettig, F. (Eds.) Attention and Vision in Language Processing. Springer.
2014	APS Student Caucus Funding Database
2012–2013	Journal of Theoretical and Artificial Intelligence

Advisor

2017	PhD, Graduate Student Mentor
2015–2017	Qualitative Measures advisor, Academic Placement Data and Analysis
2013-2015	Graduate student coordinator, Cognitive Science Student Association
2012	Co-Organizing Chair, Graduate Student Speaker Series
2010	Graduate Student Panel, Psy-Chi & SPA

Professional Memberships

Cognitive Science Society
Association for Psychological Science
American Psychological Association
Psychonomics Society
Society for Mathematics in Psychology
Society for computers in Psychology

Abstract

Intended outcomes such as expressing ideas in ways that can be understood, with the tools we know how to use, constrain the dynamics of the actions we use to ensure their success. The success of our actions can be measured by estimating the amount of information they transmit. This provides a window into the dynamics of cognition and how they are influenced by the surrounding context: Including past, present and future actions, cognitive states, social pressures and the tools we use to generate them —such as language. This dissertation is a series of studies on how context influences intended outcomes including current decisions, action dynamics and the amount of information that can be transmitted successfully. The focus is primarily on information and how it influences —and is influenced by —behavior.

The study of information cuts across disciplines. As a result, this dissertation consists of a set of interdisciplinary collaborations that quantify and explore large natural datasets. In collaboration with various coauthors from linguistics, cognitive science and applied mathematics, I present five projects that address theoretical and methodological approaches toward understanding the effects of context and what they might say about the success of our behaviors in bringing about intended effects. The studies here are presented chronologically in an effort to elucidate the process of science as it occurs naturally.

First, I present a study exploring how the Information-Theoretic structure of a message is influenced by its intended valence (Vinson & Dale, 2014b) followed by how it is influenced by social network structures (Vinson & Dale, 2016). I then report on the development of a new analysis tool that affords quantifying large text data efficiently (Vinson, Davis, Sindi, & Dale, 2016). This is followed by a study on how the dynamics of action —captured via the process of typing a message —are not encapsulated, but dynamically adjust to the statistics of the language they are used to produce. (Vinson, Dale, Shih, & Spivey, submitted). Finally, I present a study exploring how previous online business review ratings influence current ratings (Vinson, Dale, & Jones, in revision).

This dissertation, *Quantifying Context and its Effects in Large Natural Datasets*, is submitted by David W. Vinson in partial fulfillment of the degree Doctor of Philosophy in Cognitive and Information Sciences at the University of California, Merced, under the guidance of dissertation committee members Rick Dale, Teenie Matlock and Jeff Yoshimi.

Chapter 1

Introduction

> Unfortunately, the special kinds of circumstances that fit the models of a single theory turn out to be hard to find and difficult to construct. More often we must combine both knowledge and technical know-how from a larger number of different fields to product a model that will agree well enough on the matters we are looking to predict, with the method of combinations justified at best very locally. (Cartwright, 1999, p.10)

This dissertation is a collection of studies and explorations of naturally occurring data loosely driven by a much larger question; what is information and how do we share it? This question is ubiquitous across the cognitive, social, behavioral, communication and computer sciences. With the emergence of data collection techniques and machines capable of completing huge calculations efficiently, we are poised with a unique opportunity to study how information occurs naturally. As a result, the approach taken here is necessarily exploratory and best delved into through an interdisciplinary agenda.

1.1 Context

Nancy Cartwright argues that theories and discoveries do not uncover universal laws but instead *natures* (Cartwright, 1999). We discover the nature in which the phenomena observed can occur which often consists of a set of glued together theories borrowed from seemingly unrelated domains in order to make sense of what, at present, we experience. These theories, and the assumptions they hold, make up the context of the observed phenomena. But what is context? It has become an elusive placeholder used to describe any variation of implicit situational influences (Dey, 2001). Even now, modern artificial intelligence struggles to determine the boundaries of context: How much situational awareness is necessary for a machine to reason with —and abstract beyond —the current task? Yet, *context* —and our ability to harness it —is considered the next wave of machine intelligence (Launchbury, 2017).

Progress in science occurs by first defining context in an effort to ground phenomena in models of the world that afford successful actions.

> It is not the thought of the innervation which the movement requires. It is the anticipation of the movement's sensible effects, resident, or remote, and sometimes very remote indeed. Such anticipations, to say the least, determine what our movements shall be. —(James, 2013, p.521)

James is referring to the constraints that surround and influence volitional action. For instance, when we observe Jastrow's image (Figure 1.1) we can *decide* whether we will see a duck or a rabbit (Jastrow, 1900). This is not an effect of the stimuli alone, but of the cognitive context

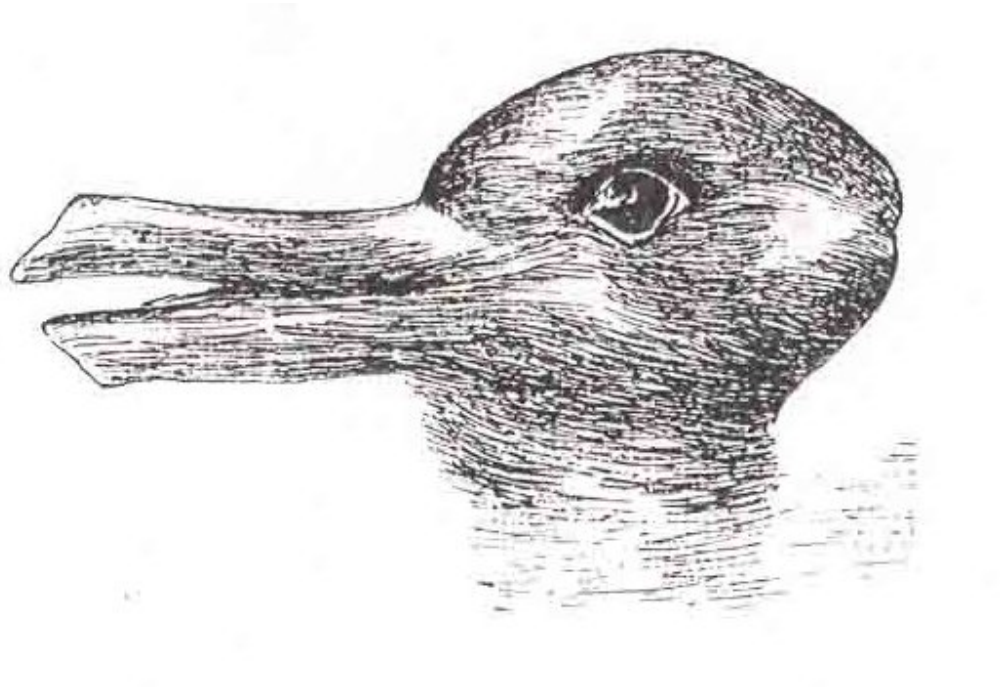


Figure 1.1: Jastrow's Duck-Rabbit Illusion

we chose to use to ground it —*rabbit or duck?*. Similarly, we define information (and in doing so ground it) in context. We operationalize both context —*I assume I will see a rabbit* —and information —*the stimuli presented appears to be a rabbit* at all levels of action. In this sense, how we make coherent sense of the world and our actions within it is, in part, driven by top-down cognitive framing effects (Vinson, Abney, et al., 2016).

The influence of context goes beyond our ability to simply choose our experiences. No behavior occurs in isolation, even those thought to be reactionary, ballistic, encapsulated and isolated from its surroundings. Recent theories suggest perceptual systems, those thought to be encapsulated due to their reflexive and automatic actions, are cognitively penetrable —influenced by higher-level cognitive phenomena such as thoughts and ideas —in so far as they lead to anticipations of incoming perceptual phenomena that reduce prediction errors (Lupyan, 2015). Indeed, anticipated outcomes occur at all levels of the cognitive system including the self sustaining dynamical nature of underlying biological procceses (Karl, 2012), neural cortical structures that support conscious thought (Dehaene, Kerszberg, & Changeux, 1998), motor action (Wolpert & Flanagan, 2001; Miall, 2003), perception (Knoblich & Jordan, 2003), attention (Simons & Chabris, 1999, for review see), agent-based behavior (A. Clark, 2013) and conscious experience (Yoshimi & Vinson, 2015). These anticipated effects, feedback projections, influence the receptive field properties of neurons throughout the visual cortex which biases vision at various levels from selective attention (Gandhi, Heeger, & Boynton, 1999; Kastner, Pinsk, De Weerd, Desimone, & Ungerleider, 1999) and contrast sensitivity (Lupyan & Spivey, 2010), to categorical decision making (Chua & Gauthier, 2015). The cognitive system *generates* predictions of incoming sensory phenomena in an effort to effectively convert sensory energy into actionable content (Lupyan, 2015). To do this, anticipated outcomes penetrate lower level sensory receptors in an effort to reduce error between the system's current state and incoming future states. This constrains everything from how we move to what we perceive. The cognitive system utilizes the underlying statistical structures of the environment to make predictions

about the world experienced next. Actions not only aim to anticipate that world, but *invent* it via the use of language and actions to modify physical and social surroundings. As a result, the dynamic nature of behaviors provides a window into contextual constraints influencing their ability to bring about intended futures.

1.2 Information: *n*-gram models of language

The study of language is a key topic in the psychological sciences. Language is influenced by a whole host of factors at many scales including vision (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), emotion (Nygaard & Queen, 2008; Pennebaker, Francis, & Booth, 2001; Pennebaker, 1997; Jurafsky, Chahuneau, Routledge, & Smith, 2014; Kahn, Tobin, Massey, & Anderson, 2007), social network structures (Vinson & Dale, 2016; Lupyan & Dale, 2010), individuals within those social networks (Nygaard, Sommers, & Pisoni, 1994; Nygaard & Pisoni, 1998; Bradlow, Nygaard, & Pisoni, 1999), gender and social status (Labov, 1972b, 1972a; Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Lindblom, 1990). Quantifying the underlying statistical distribution of language use provides a mathematical framework that can be used to determine relationships among cognitive, social and behavioral phenomena as they occur naturally (Qian & Aslin, 2014). *n*-gram models represent one of the earliest and most-used tools to uncover the statistical structure of language use. In its most basic form, an *n*-gram is a sequence of *n* items from a given collection of text, transcribed speech, genomic data, and so on (Li et al., 2001). Having originated in the early 20th century (Markov, 1913), the resurgence of *n*-gram models in language analysis came about in the mid 1970's and '80's from their successful use in speech recognition systems (Jelinek, 1976; Baker, 1975; Bahl, Jelinek, & Mercer, 1983; Martin & Jurafsky, 2000). Such systems were heavily influence by the work of Claude Shannon whose proposed theory of communication, Information Theory (Shannon, 1948), is among the most influential frameworks of the twentieth century. In fact, Shannon's key examples of Information Theory involved a *n*-gram analysis of written English. Information Theory posits that a word carries some *amount* of information, measured in *bits*, proportional to its $-\log_2$ probability of occurrence given some "context" (e.g., a corpus of words):

$$I(w_i) = -\log_2 p(w_i|w_1, w_2, ...w_{i-1}). \tag{1.1}$$

In Equation 1 above, the number of bits in word $I(w_i)$ is dependent on the frequency of its occurrence *after* some other word(s), or its maximum likelihood estimate. Though larger *n*-grams capture more linguistic structure, they are less likely to generalize across texts and even across whole corpora as the frequency of additional *n*-grams is often too small to occur more than a handful of times (Martin & Jurafsky, 2000; Piantadosi, Tily, & Gibson, 2011). Yet, the existence of larger text corpora, such as Google books and Twitter's streaming API, question whether larger *n*-gram analyses are in fact capturing rare events that only generalize to other such rare events. In the behavioral and computational sciences, this measure and related measures are very useful for exploring language production and processing.

According to our definition, for example —

> *This is a great place for lunch and dinner. The food is great, the price is good and the service is friendly and quick.*

—is a low information-dense sentence given the simplicity of the terms and their combinations. Where as —

> *This mixed media, graffiti style display withheld a deep meaning of America's religious fervor and external cultural adolescence.*

—is a high information-dense message due to the unique combination of terms used (adjusting for word length).

The amount of information that can be transferred varies across contexts and sometimes within them. This variation is a reflection of the amount of noise that exists within a communication medium. An inverse relationship exists such that as contextual noise increases, information decreases. For instance, the amount of information that can be transferred may be influenced by the language user's cognitive state, while at a more global scale, it may be influenced by the very structure of the language users social community. However, one possibility explored in Chapters 4 and 3 is how establishing common ground through increased social connectivity and shared experiences may result in a decrease in channel noise that afford information-dense language use (H. H. Clark & Brennan, 1991). Moreover, the amount of information contained within a word itself is a context that can constrain the very dynamics of behavior used to produce it (e.g., speech and typing). This notion of information extends beyond one's language use and can be used to define and quantify other behaviors. As you'll see in Chapter 6, previous actions influence current actions in ways that appear to be more closely aligned with object categorization.

1.3 Big Data and Computational Social Science

To even begin assuming we might be able to accurately capture the influence of context on language use requires huge amounts of data. The era of Big Data ushers in a new way of conceptualizing how we share and store information. Shared ideas, thoughts and interactions taking place online are captured and stored in the form of mouse movements, keystroke latencies, language use, images and more. This provides an opportunity to understand the cognitive processes underlying communication and decision making as they adapt to their context. Massive datasets provide a sea of opportunity, but also missed opportunities if the researcher does not know where to look. In an effort to harness this new sea of resources, this dissertation provides a series of ways in which the cognitive scientist can begin to successfully harness massive data sets as well as ways to collect and analyze them using theories and tools from cognitive and computer science.

Until now behavioral scientists found themselves testing specific aspects of theories in tightly controlled laboratory environments. In his paper, *Manifesto for a new (computational) cognitive revolution*, Griffiths argues that Big Data is just now making its way into the hands of behavioral scientists whose training was always geared toward showing "the value of postulating a mind between browsing history and mouse movements" (Griffiths, 2015, p.22). This dissertation harnesses natural datasets through the lens of cognitive science. Using cognitive science as a tool toward sifting through a wealth of data, we are able to develop new cognitively infused prediction models geared toward understanding behavior. In part, being able to harness these datasets involves using tools foreign to the social scientist but known to computer scientists. Not only does this demand interdisciplinary collaborations but the development of new tools.

1.4 Chapter Overview

This dissertation is formatted chronologically to elucidate the process of science. As detailed above, the work here is highly theoretically motivated (*chapters 2, 4, 5 & 6*), but in practice scientists must occasionally address problems at the boundaries of their own knowledge, and often tertiary to progress in areas of theoretical interest. In many cases, solving these problems is the result of successful interdisciplinary collaboration. *Chapter 3* is the result of one successful collaboration between the cognitive and information sciences and applied

Table 1.1: Context operationalized by Chapter

Chapter	Context	Operationalized
All	Language Use	Information Theory
2	Cognitive State	Review Ratings
3	Community	Social Network Analysis
4	Audience	Useful|Funny|Cool Ratings
5	Action Dynamics	Typing Rate
6	Past Actions	Prior Review Ratings

mathematics groups at the University of California, Merced. Table 1 highlights how context is operationalized in each chapter.

In Chapter 2, we sought a structured corpus in order to extend and explore the potential role of context in the observed information density of messages. This was a first attempt at harnessing large natural behavior datasets. We used a database of over one hundred thousand consumer reviews that includes an assortment of user-related variables. These user-related variables, such as the overall rating of a review, appear to have an interesting relationship to basic information-theoretic measures, such as the average amount and variability of observed information of a review's words. We discuss these results in terms of the broader context that may shape the information structure of messages and relate these findings to existing theories.

In chapter 3, we investigate how language adapts to changes in the structure of its social community. Some recent theories of language see it as a complex and highly adaptive system, adjusting to factors at various time scales. For example, at a longer time scale, language may adapt to certain social or demographic variables of a linguistic community. At a shorter time scale, patterns of language use may be adjusted by social structures in real-time. Until recently, datasets large enough to test how socio-cultural properties— spanning vast amounts of time and space— influence language change have been difficult to obtain. The emergence of digital computing and storage have brought about an unprecedented ability to collect and classify massive amounts of data. By harnessing the power of big data, we can explore what socio-cultural properties influence language use. This chapter explores how social-network structures, in general, contribute to differences in language use. We analyzed over one million online business reviews using network analyses and information theory to quantify social connectivity and language use. Results indicate that perhaps a surprising proportion of variance in individual language use can be accounted for by subtle differences in social-network structures, even after fairly aggressive covariates have been added to regression models. The benefits of utilizing big data as a tool for testing classic theories in cognitive science and as a method toward guiding future research are discussed.

Cognitive and social science theories and questions within the emerging environment of big data can result in the need for practical problems to be solved before additional progress can be made. That is, the computational social scientist is often faced with a need to either format or constrain their questions in ways that can be solved using well-known behavioral/-experimental methods, or reformat their tools in ways that can address the question of interest. After our initial investigation into how cognitive and social effects influence the structure of language use, we ran into computational bottlenecks when analyzing reviews. Specifically, the dataset we were using increased 10 fold. In Chapter 4 we address this by introducing a new R package `cmscu`, which implements a Count-Min-Sketch with conservative updating (Cormode & Muthukrishnan, 2005), and its application to n-gram analyses (Goyal, Daumé III, & Cormode, 2012). By writing the core implementation in C++ and exposing it to R via Rcpp, we are able to provide a memory-efficient, high-throughput, and easy-to-use library. As a proof of concept, we implemented the computationally challenging (Heafield, Pouzyrevsky, Clark, &

Koehn, 2013) modified Kneser-Ney n-gram smoothing algorithm using `cmscu` as the querying engine. We then explore information density measures (Jaeger, 2010) from n-gram frequencies (for $n = 2,3$) derived from a corpus of over 2.2 million reviews provided by a Yelp, Inc. dataset. We demonstrate that these text data are at a scale beyond the reach of other more common, more general-purpose libraries available through CRAN. Using the `cmscu` library and the smoothing implementation, we find a positive relationship between review information density and reader review ratings. We end by highlighting the important use of new efficient tools to explore behavioral phenomena in large, relatively noisy data sets.

In Chapter 5, we take a different look at information. Instead of treating it as the dependent variable —testing what contextual effects might influence it, we consider information the context and see whether the dynamics of micro-behaviors are constrained by this statistical context. We show that the microstructure of language production —expressed in the keystroke-to-keystroke dynamics of spontaneous writing —shows a reliable "echo" of word predictability. At the keystroke level, there is an approximate logarithmic relationship between word predictability and typing speed. In many prior studies of typing, such an effect was difficult to obtain. A unique feature of the data shown here is that it is based on a large dataset of spontaneous language production, obtained via free composition through typing. We show, and replicate, curious effects of probabilistic structure. When conditioning predictability of a word by its preceding word, this conditional probability covaries most strongly with the speed of function words. When considering the predictability of a word from raw frequency, this more strongly covaries with content words. In sum, when combined with corpus estimations of word predictability, it may be possible to capture "echoes" of the flow of the cognitive processes that support language production, at least in the composition of text. This challenges long-standing theories of the encapsulation of the production system during typing, even in highly practiced words: These effects hold even for the English determiner "the."

Finally, in Chapter 6, we take the underlying assumptions of information theory and use it to determine if there is an influence of previous actions on current behavior. This chapter provides evidence that past behavior not only influences current behavior but may make up some portion of the context the current behavior is nested within. Current judgments are systematically biased by prior judgments. Such biases occur in ways that seem to reflect the cognitive system's ability to adapt to statistical regularities within the environment. These cognitive sequential dependencies have primarily been evaluated in carefully controlled laboratory experiments. The study uses well-known laboratory findings to guide the analysis of two datasets consisting of over 2.2 million business review ratings from Yelp and 4.2 million movie and television review ratings from Amazon. The study explores how within-reviewer ratings are influenced by previous ratings. Findings suggest a contrast effect: Current ratings are systematically biased away from prior ratings, and the magnitude of this bias decays over several reviews. Prior reviews "contaminate" future reviews showing weak but reliable "echoes" of previous decisions influencing current ones. current ones. principle influencing a natural dataset. This work is couched within a broader program that aims to use well-established laboratory findings to guide our understanding of patterns in naturally occurring and large-scale behavioral data.

Throughout this work a few themes are present including information theory, big data, social networks, decision making and their practical applications. Further, the successful transfer of ideas from one agent to another is highly dependent on its context. This theme is ubiquitous throughout the following work. Chapter 7 concludes by discussing how these studies and themes fit together into a coherent framework, via the cognitive system's use of the underlying statistical structure within its environment to predict and create future events. Future directions are also discussed. This dissertation contributes to the fields of cognitive science, statistics, linguistics, social science, computer science that together form the emerging field of computational social science.

Chapter 2

Valence Weakly Constrains the Information Density of Messages

2.1 Introduction

Tools from information theory have allowed researchers to explore whether language use is, in some sense, optimal, e.g., (Jaeger & Levy, 2007). At the production level, speakers may structure their utterances so as to optimize information density (Jaeger, 2010), while over longer timescales aspects of language such as word length, may be optimized according to information content (Piantadosi et al., 2011).

In most cases, factors beyond the lexical level that may influence information density must be abstracted away. For example, "context" is often confined to a lexical definition, namely the immediate preceding word. In this case, the information encoded by a word can be expressed using the log of the probability that the word would occur in this lexical context:

Though easy to compute, this definition abstracts away a variety of other contextual factors that may help explain why a user chooses a given word. This simplification is justifiable, of course, because of the difficulty in defining other contextual factors (e.g., at a semantic level), and the complexity that seems endemic to high-level aspects of language (see Jaeger, 2010 for discussion).

More recently, studies have begun to show that information density is influenced by factors at a variety of linguistic levels including syntactic variation and phonetic reduction (Aylett & Turk, 2004; Jaeger, 2010; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013). Relatedly, the information density of a linguistic message may be subject to more social or cognitive constraints that help define the content of a message, reaching beyond local phonological and syntactic levels. In other words, in abstract terms, the transfer of information may be subject to a variety of ubiquitous contextual constraints, at a variety of levels.

One such constraint, and one of interest to the current paper, is the relative valence of a linguistic message (and potentially, the language user herself). If a message is intended to be a highly positive evaluation of some situation, does the language user seek different patterns of information density to convey it?

As we review below, there is reason to suspect that such a pervasive contextual variable —one specific to the user while composing a message —may shape the information content of that message. Evidence for this would further support the idea that the information-theoretic properties of language are contextually modulated. We sought a corpus well suited to test this idea. We analyzed over $100,000$ consumer reviews with associated information about a review's rating. Even after accounting for a variety other linguistic variables, findings show affective states influence the amount of information transmitted. Crucially, specific findings depend on

how information is quantified.

2.1.1 Lexical Constraints on Information

Studies that stem from an information-theoretic standpoint have only recently begun to theorize what contextual effects influence the transfer of information at a lexical level (Aylett, 1999; Genzel & Charniak, 2002). One such theory, known as Uniform Information Density (UID), states that language users will structure their utterances so as to optimize information transfer within a given context (Frank & Jaeger, 2008; Jaeger & Levy, 2007). That is, a speaker will communicate at a rate that is optimal for transferring the greatest amount of information within a specific (noisy) channel, without loss of information or miscommunication (Genzel & Charniak, 2002). Recent evidence supporting this notion shows speakers may be sensitive to linguistic probability distributions that help define the information density of a message (Fine, Qian, Jaeger, & Jacobs, 2010; Fine & Jaeger, 2011).

In support of optimal information transfer theories Aylett (1999) found individuals take longer to communicate more information dense messages. In addition, Jaeger (2010) found that speakers' use of an optional "that" complementizer is dependent on the information density of their utterance. Further, Piantadosi, Tilly and Gibson (2011) show word length in general may be optimized to the amount of information transferred, in contrast to the well known Zipf's law —word length is optimized for the frequency of word use (Zipf, 1949). Each study shows information density optimization occurs in subtly different, but related ways.

Crucially, a word's lexical context stands as the primary constraint guiding one's understanding of the amount of information that can be transmitted by any given message; even though other, higher-level visual and social constraints are known to influence language use and comprehension (Tanenhaus et al., 1995; Vinson, Dale, Tabatabaeian, & Duran, 2015). Importantly, if individuals are sensitive to specific linguistic distributions, social or cognitive factors influencing such distributions may affect the density of information within a message.

2.1.2 Affect and Message Valence

The information density of an message is at least partially dependent on contextual constraints such as the current lexical context. However, this lexical context may be further influenced by other, more global contextual constraints.

Several findings, especially in social cognition, recommend this hypothesis. In particular, past research has suggested that a cognitive or affective state with more positive valence is likely to generate more flexible, open-ended behaviors (Cacioppo & Gardner, 1999; Diener & Diener, 1996; Fredrickson, 2004; Isen & Means, 1983). Consider an example study that shows this tendency. Positive affective states in doctors lead to the correct diagnosis for patient symptoms quicker compared to doctors not primed to experience a positive affective state (Estrada, Isen, & Young, 1997). Doctors were more likely to accept new information when in a positive affective state than when in a neutral state. Similarly, it may be that when experiencing a positive affective state, one might transmit a more information dense message than when in a less positive state.

This notion finds relevance in information-theoretic terms where positive valence may provide an appropriate context for transferring more information. We speculate further on this relationship below, but one possibility is that particular cognitive-affective states may increase the channel capacity for both sender and recipient. Though it is a provocative hypothesis, the corpus we use here provides a massive amount of text data where individuals label their experiences as positive or negative on a scale of 1-5 and briefly report on them. We speculate that one's experiential rating provides a measures whereby the influence of one's affect valence on information density can be assessed; even if only weakly connected.

We hypothesize that when individuals experience a positive affective state, their use of language may be more informative, more lexically rich and differ in frequency of use compared to individuals in a more negative affective state. Provided this hypothesis, one's affective state, may be predicted by their language use; acting as a constraint on the density of information transferred over the course of a single message or, in this case, a business review.

2.2 Current Study

The current study used a dataset from Yelp, Inc. consisting of over 100,000 consumer reviews of businesses throughout the city of Phoenix AZ. This dataset consists of written reviews associated with the reviewer's explicit feelings specified by a rating from 1 (negative) to 5 (positive) stars about the business reviewed. Each review was subject to being labeled useful, funny, or cool by other reviewers. For the purpose of this study we assume the cognitive-affective state of an individual is in some way correlated to the number of stars associated with their review; more stars being correlated with a more positive affective state while fewer stars are correlated with a negative affective state. Because this corpus of reviews consists of both an explicit rating of the business and a linguistic message about the consumer's experience, it is highly suitable to testing how various contextual factors, in particular cognitive-affective states, influence the information density of a linguistic message.

2.2.1 Measures and Method

Prior to testing how a reviewer's cognitive state might influence the information density of their message, we must define measures of information that seem relevant. This has been done in a variety of ways. Here we define information in four very simple ways, commensurate with classic information-theoretic definitions. Each function defines the linguistic context of an utterance slightly differently. Importantly, such differences might reveal a unique relationship to valence. Listed here are the four functions along with a brief definition of each:

(1) Review-internal entropy (RI-Ent). A review may simply be structured in distinct ways depending on how a language user decides to use lexical tokens more or less regularly in a way purely internally to a review itself. In other words, the frequency distribution over words may reflect diverse selection of types (higher entropy), or it may be relatively more repetitive (lower entropy). This can be expressed in the following way:

$$\textit{RI-Ent}_j = -\sum_{i=1}^{N} p(w_i|R_j) log_2 p(w_i|R_j) \tag{2.1}$$

Here, $RI-Ent_j$ denotes the j^{th} review, containing N words, as the probability of the i^{th} word occurring within that review (for notational convenience we treat it as a conditional probability, equivalent to restricting computations to a given review). This measure can be seen as a kind of lexical richness score, expressed as the expected number of bits required to encode a message, given its unique internal word distribution. If information density is high, the text can be said to be lexically rich. Indeed, it can be easily shown that RI-Ent correlates with common measures of lexical richness, such as type-token frequency. Put simply, a review with higher entropy will have more unique tokens, thus being, in a sense, more "information dense."

(2) Average unigram information (AUI). This measure is computed from the lexical distribution over the entire set of Yelp reviews. As noted in the introduction, the information encoded in a word can be simply seen as the negative log of the probability of its occurrence (the less probable a word, the more informative). For any given review j:

$$AUI_j = -\frac{1}{N}\sum_{i=1}^{N} log_2 p(w_i) \tag{2.2}$$

This differs from the previous measure in that the probability of a word's occurrence is defined by a much larger distribution of words. If we regard the overall distribution of terms in Yelp as a simple but direct measure of how informative a word is, then a review may vary in its informational content depending on the language user's state.

(3) Average conditional information (ACI). A more common way of expressing the information encoded in a word is relative to some context (i.e., a second-order estimate). As noted in the introduction, this is commonly taken to be some immediate lexical context. In our case, we extract a very simple contextual information measure:

$$ACI_j = -\frac{1}{N-1}\sum_{i=2}^{N} log_2 p(w_i|w_{i-1}) \tag{2.3}$$

Here the information in a word is the negative log of the probability of its occurrence given the previous word. This differs from RI-Ent and AUI in that it accounts for the most immediate or local context, namely, the previous word.

(4) Conditional information variability (CIV). ACI reflects the average information, but the work of Jaeger (2010) and Levy and Jaeger (2007) suggest that the uniformity, or variability, of this information measure may be interesting to explore.

$$CIV_j = \sigma(CI_j) \tag{2.4}$$

Here, CI_j is the set of conditional information scores for each word of the j^{th} review; we compute the standard deviation of this set. Greater variability in information density would reflect an increase in the channel capacity. This would permit more variability in word choice allowing differences in the rate of information transmitted (i.e., by diminishing range restriction).

From these measures, reviews can be defined as more or less information dense depending on both their general and local linguistic contexts. Distributions of these measures over more than $100,000$ reviews are shown in with two example reviews in Figure 2.1. Using simple measures we tested if information encoded in a message is related to cognitive context: the intended valence of that message. To test this, we use star rating to predict information in regression models: Does variation in valence (rating) predict the level of information encoded.

$124,622$ Yelp reviews [1] were imported and processed in Python using json. We used nltk and numpy/scipy libraries to carry out most calculations. To calculate $RI-Ent$ we used nltk's MLE entropy function.

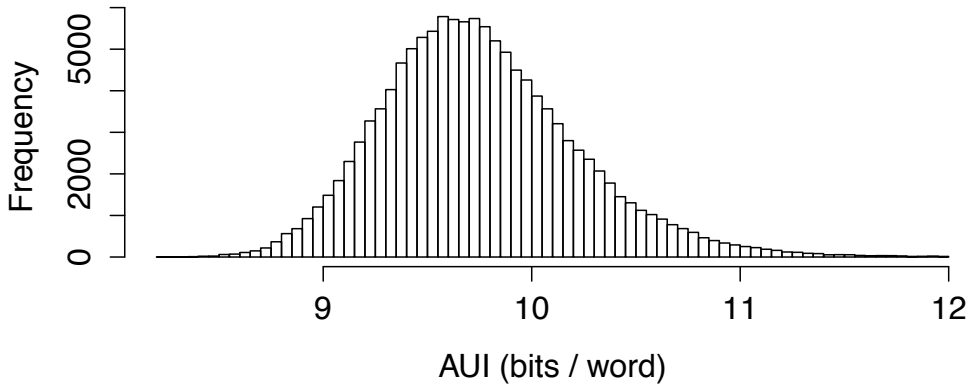
2.3 Results

At least one obvious measure may correlate with star ratings: review length (in number of words). We first test this variable and then include it as a covariate when testing our key information-theoretic measures (see Figure 2.2).

Review length. It is well known that bin size can impact our key information-theoretic measures. In fact, review length indeed differed by star rating. We used a simple linear model to predict review length by stars.[2] There were significantly more words per review for lower stars ($R^2 = .01$, $t(124,621) = -38.7$, $p < .001$). This represents a small but significant effect —detectable thanks to the massive power of the large Yelp corpus. We used review length as a covariate in our subsequent analyses of information-theoretic measures.

[1] The full Yelp dataset contains about $229,000$ reviews. We filtered this dataset by choosing reviews with 100 words or more so as to increase the reliability of our information measures.

[2] Due to the computation required in estimating models from so much data, we chose simple and multiple regression with lm in R; we also confirmed general patterns by centering scores relative to reviewers, and exploring linear mixed-effects models.



Very low AUI: “This is a great place for lunch and dinner. The food is great, the price is good and the service is friendly and quick.” [AUI = 7.4]

Very high AUI: “I don't know if this qualifies as an update. However 101 Bistro is now closed. Eighty sixed. Nada here anymora. Adios. Hasta la pasta.” [AUI = 13.9]

Figure 2.1: Distribution of Average Unigram Information across Yelp Reviews

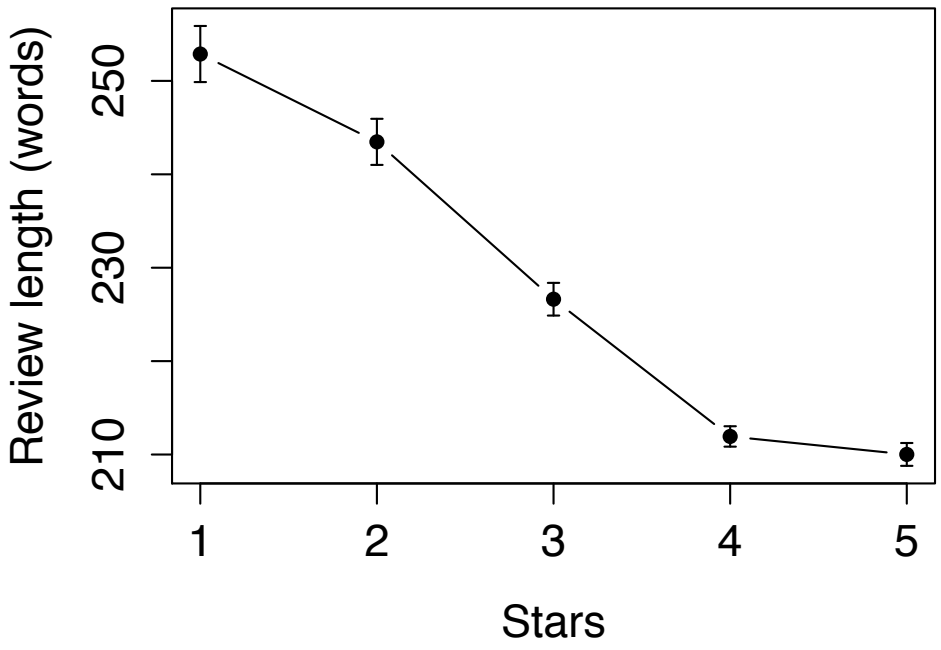


Figure 2.2: Review Length (y-axis) by Star Rating (x-axis) with 95% Confidence Intervals over the filtered Yelp dataset

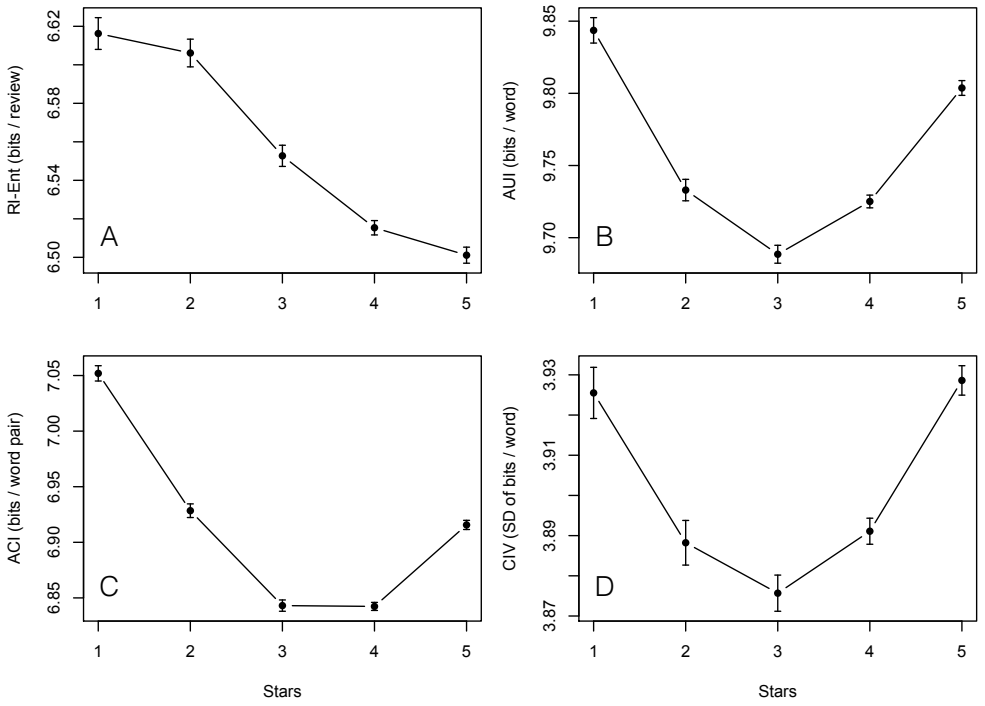


Figure 2.3: Initial relationships, without additional covariates, between Star Ratings and (A) Review-Internal Entropy ($RI-Ent$), (B) Average Unigram Information (AUI), (C) Average Conditional Information (ACI), and (D) Conditional Information Variability (CIV)

(1) Review-internal entropy (RI-Ent). When not controlling for review length there was a small, but highly significant effect of stars in predicting $RI-Ent$ ($R^2 = .009$, $t(124,621) = -34.01$, $p < .001$). Again, this shows a reliable but weak effect; stars account significantly for about 1% of the variance in $RI-Ent$ (see Figure 2.3A).

We controlled for review length by fully residualizing $RI-Ent$ in the following way: We predicted $RI-Ent$ by review length, and stored the residuals as a new outcome variable for the linear model with stars as the predictor. Residuals would therefore reflect unique variance associated with star rating in predicting $RI-Ent$. When doing this, there is no longer a significant effect of rating ($R^2 = 0$, $t(124,621) = -0.43$, $p = .67$). It appears that the variability present in $RI-Ent$ does not covary with affective state when review length is controlled.

(2) Average unigram information (AUI). Interestingly, and unexpectedly, AUI shows a quadratic relationship with stars (Hu, Pavlou, & Zhang, 2006). This is plainly seen in see Figure 2.3B. To model this, we converted stars into a quadratic term ($[1,2,3,4,5] = [4,1,0,1,4]$). When not controlling for review length the raw analysis revealed a small but highly significant effect of stars in predicting AUI ($R^2 = .010$, $t(124,621) = 36.24$, $p < .001$), such that information density of a review increased as rating levels became more extreme.

When taking out review length, and running the regression with residuals, this effect remained ($R^2 = .011$, $t(124,621) = 36.90$, $p < .001$), suggesting it is highly independent of review length. Again, though a weak effect, there is a relationship between the average single-word information of reviews and star ratings. It appears that positive valence is not predictive of overall information; rather, the extremity of the valence predicts slightly more loading of

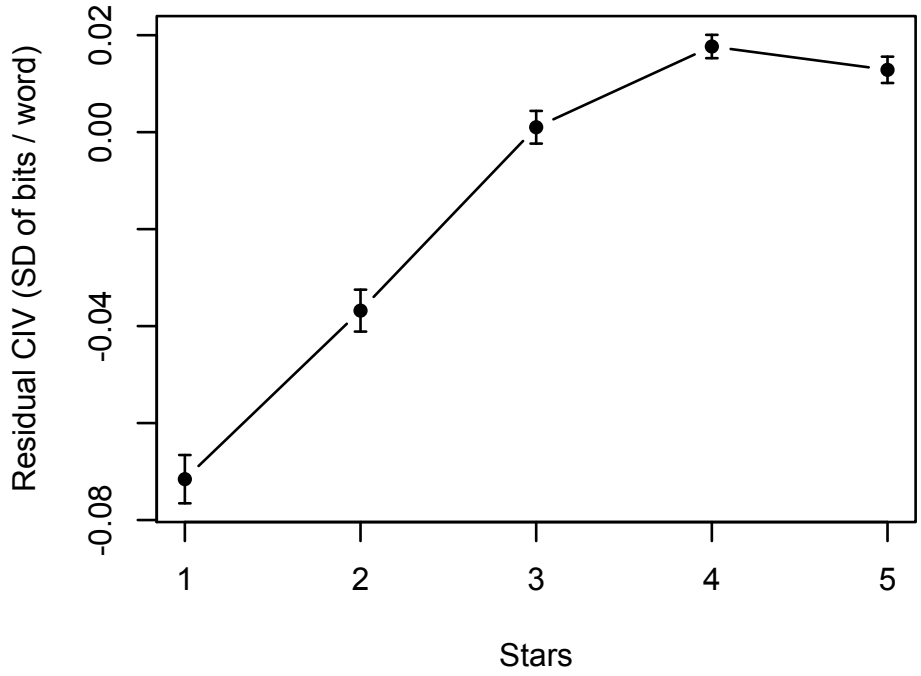


Figure 2.4: CIV and Star Ratings

reviews with greater information density (i.e., equivalently, lower frequency terms).

(3) Average conditional information (ACI). Interestingly, *ACI* also showed a nonlinear relationship with stars, shown in Figure 2.3C. With star rating predicting *ACI* alone, there is a significant quadratic relationship ($R^2 = .013$, $t(124,621) = 39.89$, $p < .001$). The same pattern appears to hold: Extreme reviews seem to generate more information in bigrams patterns, though this seems to be more pronounced in the negative reviews. When residualizing *ACI* by review length and AUI (as an additional covariate), this relationship shrinks in effect, but remains statistically reliable ($R^2 = .003$, $t(124,621) = 20.01$, $p < .001$).

(4) Conditional information variability (CIV). At first blush, *CIV* also has a nonlinear relationship with stars (Figure 2.3D). Again, the quadratic term for stars significantly predicts *CIV* scores ($R^2 = .004$, $t(124,621) = 21.35$, $p < .001$), though the effect is even smaller. However, we controlled for both review length and *ACI*, since the height of *ACI* will generally correlate with *CIV*'s range (due to range restriction with a true 0). When we do this, the relationship between *CIV* and stars completely changes (see Figure 2.4). Now, in fact, greater informational variability appears to be related to increased positive valence. This relationship, between the residual of *CIV* and star rating, is statistically reliable but again very small ($R^2 = .009$, $t(124,621) = 32.86$, $p < .001$).

Other user-related variables. The Yelp dataset also allows us to explore information as it relates to the listener" in this context. Users who read reviews have the option to rate them as useful, cool, or funny. Exploration with simple logistic regression finds that even using these simple, surface information-theoretic measures provides a boost in predicting whether a review will be categorized as fun or cool (useful is not predicted by information measures, surprisingly). Some details are shown in Table 2.1, detailing fully specified models with centered interaction terms for all information-theory values, review length, and comparison models.

Table 2.1: Basic results of logistic regression when categorizing reviews along certain "listener" dimensions.

	Full Model	*Length Only*	*Intercept Only*
Cool?	57.9% (168117)	55.8% (170419)	52.7% (172411)
Useful?	68.5% (152015)	68.5% (152976)	68.5% (155378)
Funny?	63.9% (159643)	62.4% (163005)	61.4% (166277)

Note: Categorization uses a 0.5 threshold in *GLM* predictions using `family=binomial(logit)`. *AIC* shown in parentheses. All full models have lower *AIC*, though performance difference is small.

2.4 General Discussion

Variance in information density is partially, if only weakly, captured by the review's star rating. Though we obtain very small effects overall, we would argue that these remain theoretically intriguing. For example, we find a curious and unpredicted quadratic relationship between average lexical information and review rating. This suggests participants may be choosing lower frequent terms "greater lexical richness" when composing reviews at the extremes of the scale (in contrast to our hypothesis that positive reviews, specifically, would be of greater lexical richness).

Overall, results suggest variability in the review information density is at least partially accounted for by contextual influences beyond a linguistic level. If star rating is an indication of a reviewer's affective valence, then the cognitive state of a reviewer may stand as one contextual factor that can account for changes in the information density expressed in a message. This provides further evidence that information density is in general dependent on contextual factors such as efficiency (Jaeger, 2010; Genzel & Charniak, 2002) and possibly affective valence. If this is true, it may be that speakers are sensitive to a variety of linguistic probability distributions well suited to convey one's message under any variety of constraints. Perhaps it should be no surprise that the content of a message expressing intense joy is information-theoretically different, slightly, from one of mediocrity.

An intriguing pattern of effects is seen in the *CIV* measure. When controlling for *ACI*, results completely changed, yet remained highly significant such that some amount of variance in *CIV* was accounted for by star rating (i.e., increasing as star rating increased). As a general principle underlying uniform information density, when information variability increases the rate of information transfer may be less uniform. This suggests the affective valence of a message, or the cognitive state of a speaker for that matter, might moderate how much information can be transferred within a message. This suggests that what is considered optimal under principles of optimal information transfer, may be subject to a variety of higher-level constraints (see (Ferrer-i Cancho, Debowski, & del Prado Martín, 2013; Mahowald et al., 2013) for a more recent debate over the use of constant entropy rate measures in describing language use).

In summary, speakers may be sensitive to the rate of information in sequencing their message, adjusting their message according to a particular rate of information transfer (Jaeger, 2010; Fine et al., 2010). Our results are commensurate, in a way, with this intuition. The information density of a message, and the variability of that density are sensitive, at least weakly, to the message's affective valence. This provides further support for the idea that a speaker's word choice is subject to a variety of ubiquitous, and perhaps multi-scaled, constraints.

Chapter 3

Social Structure Relates to Linguistic Information Density

3.1 Introduction

Language is a complex behavioral repertoire in a cognitively advanced species. The sounds, words, and syntactic patterns of language vary quite widely across human groups, who have developed different linguistic patterns over a long stretch of time and physical separation (Sapir, 2004). Explanations for this variation derive from two very different traditions. In the first, many language scientists have sought to abstract away from this observed variability to discern core characteristics of language which are universal and perhaps genetically fixed across people (Chomsky, 1975; Hauser, Chomsky, & Fitch, 2002). The second tradition sees variability as the mark of an intrinsically adaptive system (Christiansen & Chater, 2008). For example, Beckner and colleagues (2009) argue that language should be treated as responsive to socio-cultural change in real time. Instead of abstracting away apparently superficial variability in languages, this variability may be an echo of pervasive adaptation, from subtle modulation of real-time language use, to substantial linguistic change over longer stretches of time. This second tradition places language in the broader sphere of human behavior and cultural products in a time when environmental constraints have well known effects on many aspects of human behavior (Triandis, 1994).[1]

Given these explanatory tendencies, theorists of language can have starkly divergent ideas of it. An important next step in theoretical mitigation will be new tools and broad data samples so that, perhaps at last, analyses can match theory in extent and significance. Before the arrival of modern information technologies, a sufficient linguistic corpus would have taken years, if not an entire lifetime, to acquire. Indeed, some projects on the topic of linguistic diversity have this property of impressive timescale and rigor. Some examples include the Philadelphia Neighborhood Corpus, compiled by William Labov in the early 1970s, the Ethnologue, first compiled by Richard Pittman dating back to the early 1950's and the World Atlas of Language Structures (WALS) (Dryer et al., 2005), a collection of data and research from 55 authors on language structures available online, only six years ago (in 2008). Digitally stored language, and to a great extent accessible for analysis, has begun to exceed several exabytes,

[1] This description involves some convenient simplification. Some abstract and genetic notions of language also embrace ideas of adaptation (Pinker & Bloom, 1990), and other sources of theoretical subtlety render our description of the two traditions an admittedly expository approximation. However, the distinction between these traditions is stark enough to warrant the approximation: The adaptive approach sees all levels of language as adaptive across multiple time scales, whereas more fixed, abstract notions of language see it as only adaptive in a restricted range of linguistic characteristics.

generated everyday online (Kubyba & Kwatinetz, 2014).[2] One way this profound new capability can be harnessed is by recasting current theoretical foundations, generalized from earlier small-scale laboratory studies, into a big-data framework.

If language is pervasively adaptive, and is thus shaped by socio-cultural constraints, then this influence must be acting somehow in the present day, in real-time language use. Broader linguistic diversity and its socio-cultural factors reflect a culmination of many smaller, local changes in the incremental choices of language users. These local changes would likely be quite small, and not easily discerned by simple observation, and certainly not without massive amounts of data. In this chapter, we use a large source of language data, Yelp, Inc. business reviews, to test whether social-network structures relate in systematic ways to the language used in these reviews. We frame social-network variables in terms of well-known network measures, such as centrality and transitivity (Bullmore & Sporns, 2009), and relate these measures to language measures derived from information theory, such as information density and uniformity (Aylett, 1999; Jaeger, 2010; Jaeger & Levy, 2007) (Aylett, 1999). In general, we find subtle but detectable relationships between these two groups of variables. In what follows, we first motivate the broad theoretical framing of our big-data question: What shapes linguistic diversity and language change in the broad historical context? Following this we describe information theory and its use in quantifying language use. Then, we explain how social structure may influence language structure. We consider this a first step in understanding how theories in cognitive and computational social science can be used to harness the power of big data in important and meaningful ways (Griffiths, 2015).

3.1.1 What shapes language?

As described above, language can be cast as a highly adaptive behavioral property. If so, we would probably look to social, cultural or even ecological aspects of the environment to understand how it changes (Nettle, 1998; Nichols, 1992; Trudgill, 1989, 2011). Many studies, most over the past decade, suggest languages are dynamically constrained by a diverse range of environmental factors. Differences in the spread and density of language use (Lupyan & Dale, 2010), the ability of its users (Bentz, Verkerk, Kiela, Hill, & Buttery, 2015; Chater, Reali, & Christiansen, 2009; Dale & Lupyan, 2012; Ramscar, 2013; Wray & Grace, 2007) and its physical environment (Ember & Ember, 2007; Everett, 2013; Nettle, 1998) impact how a language is shaped online (Labov, 1972a, 1972b) and over time (Nowak, Komarova, Niyogi, et al., 2002). These factors determine whether certain aspects of a language will persist or die (Abrams & Strogatz, 2003), simplify or remain complex (Lieberman, Michel, Jackson, Tang, & Nowak, 2007). Language change is also rapid, accelerating at a rate closer to that of the spread of agriculture (Chater et al., 2009; Gray & Atkinson, 2003) than genetics. Using data recently made available from WALS and a recent version of the Ethnologue (Gordon, Grimes, et al., 2005), Lupyan and Dale (2010) found larger populations of speakers, spread over a wider geographical space, use less inflection and more lexical devices. This difference may be due to differences in communicating within smaller, "esoteric" niches and larger, "exoteric" niches (Wray & Grace, 2007), such as the ability of its speakers (Bentz & Winter, 2013; Dale & Lupyan, 2012; Lupyan & Dale, 2010) or one's exposure to a growing vocabulary (McWhorter, 2002; Reali, Chater, & Christiansen, 2014).

Further evidence of socio-cultural effects may be present in real-time language usage. This is a goal of the current chapter — can we detect these population-level effects in a large database of language use? Before describing our study, we describe two key motivations of our proposed analyses: The useful application of (1) information theory in quantifying language use

[2] Massive online sites capable of collecting terabytes of metadata per day have only emerged in the last 10 years: Google started in 1998; Myspace 2003; Facebook, 2004; Yelp 2004; Google+ 2011. Volume, velocity and variety of incoming data are thought to be the biggest challenges toward understanding big-data today (McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012).

and (2) network theory in quantifying social structure.

3.1.2 Information and adaptation

Information theory (Shannon, 1948) defines the second-order information of a word as the negative log probability of a word occurring after some other word:

$$I(w_i) = -log_2 p(w_i|w_{i-1}) \tag{3.1}$$

The theory of uniform information density; `UID` (Jaeger, 2010; Jaeger & Levy, 2007) states that speakers will aim to present the highest amount of information across a message at a uniform rate, so as to efficiently communicate the most content without violating a comprehender's channel capacity. Support for this theory comes from Aylett (1999), in an early expression of this account, who found that speech is slower when a message is informationally dense and (Jaeger, 2010), who found information-dense messages are more susceptible to optional word injections, diluting its overall density over time. Indeed, even word length may be adapted for its informational qualities (Piantadosi et al., 2011).

In a recent paper, we investigated how a simple contextual influence, the intended valence of a message, influences information density and uniformity. While it is obvious that positive and negative emotions influence what words individuals use (Vinson & Dale, 2014a), it is less obvious that the probability structure of language use is also influenced by one's intentions. Using a corpus of over two-hundred thousand online customer business reviews from Yelp, Inc., findings suggest that the information density of a message increases as the valence of that message becomes more extreme (positive or negative). It also becomes more uniform (less variable) as message valence becomes more positive (Vinson & Dale, 2014b). The results are commensurate with theories that suggest language use adapts to a variety of socio-cultural factors in real time. In this chapter, we look to information-theoretic measures of these kinds to quantify aspects of language use, with the expectation that they will also relate in interesting ways to social structure.

3.1.3 Social network structure

Another key motivation of our proposed analyses involves the use of network theory to quantify the intricate structural properties that connect a community of speakers (Christakis & Fowler, 2009; Lazer et al., 2009). Understanding how specific socio-cultural properties influence language can provide insight into the behavior of the language user herself (Baronchelli, Ferrer-i Cancho, Pastor-Satorras, Chater, & Christiansen, 2013). For instance, Kramer, Guilliory and Hancock (2014) having analyzed over six hundred thousand Facebook users, reported when a user's newsfeed was manipulated to show only those posts that were either positive or negative, a reader's own posts aligned with the emotional valence of their friends' messages. Understanding what a language looks like when under certain socio-cultural pressures can provide valuable insight into what societal pressures that help shape a language. Indeed, global changes to one's socio-cultural context, such as changes in the classification of severity of crime and punishment over time, are marked by linguistic change (Klingenstein, Hitchcock, & DeDeo, 2014) while differences in the distance between socio-cultural niches is marked by differences in language use (Vilhena et al., 2014).

3.2 Current Study

In the current study, we utilize the Yelp database as an arena to test how population-level differences might relate to language use. While previous work suggests online business reviews may provide insight into the psychological states of its individual reviewers (Jurafsky

Table 3.1: Summary of Information-Theoretic measures

Measure	Description	Definition
$RI-Ent$	Review-Internal Entropy	$\textit{RI-Ent}_j = -\sum_{i=1}^{N} p(w_i\|R_j) log_2 p(w_i\|R_j)$
AUI	Average Unigram Information	$AUI_j = -\frac{1}{N}\sum_{i=1}^{N} log_2 p(w_i)$
ACI	Average Conditional Information	$ACI_j = -\frac{1}{N-1}\sum_{i=2}^{N} log_2 p(w_i\|w_{i-1})$
UIV	Unigram Information Variability	$UIV_j = \sigma(UI_j)$
CIV	Conditional Information Variability	$CIV_j = \sigma(CI_j)$

Note: N = number of words in a review; $p(w)$ = probability of word w; w_i = ith word of a review; UI_j = set of unigram information scores for each word of a given review; CI_j = set of conditional information scores for each word of a given review.

et al., 2014), we expect that structural differences in one's social community as a whole, where language is crucial to conveying ideas, will affect language use. We focus on how a language user's social niche influences the amount and rate of information transferred across a message. Agent-based simulations (Christiansen, Reali, & Chater, 2006; Dale & Lupyan, 2012; Reali et al., 2014) and recent studies on the influences of interaction in social networks (Choi, Blumen, Congleton, & Rajaram, 2014)(Bond et al., 2012) indicate that the structure of language use may be influenced by structural aspects of a language user's social interactions. From an exploratory standpoint, we aim to determine if one's social-network structure predicts the probability structure of language use.

3.3 Methods

3.3.1 Corpus

We used the Yelp Challenge Dataset[3]. At the time of this analysis the dataset contained reviews from businesses in Phoenix, Las Vegas, Madison, and Edinburgh. This includes 1,125,458 reviews from 252,898 users who reviewed businesses in these cities. The field entries for reviews included almost all the information that is supplied on the Yelp website itself, including the content of the review, whether the review was useful or funny, the star rating that was conferred upon the business, and so on. It omits a user's public username, but includes an array of other useful information, in particular a list of user ID codes that point to friends of a given user. Yelp users are free to ask any other Yelp user to be their friend. Friendship connections are driven by users' mutual agreement to become friends. These user ID codes allow us to iteratively build social networks by randomly choosing a user, and expanding the network by connecting friends and friends of friends, which we further detail below.

3.3.2 Linguistic measures

The first and simplest measure we explore in our analysis is the number of words in a given review, its word length. This surface measure is used as a basic but important covariate for regression analyses. Word length will define the bin count for entropy and other information analyses, and so directly impacts these measures.

The second measure we use is the reviewer-internal entropy ($RI-Ent$) of a reviewer's word use. This marks the discrete Shannon entropy of a reviewer's overall word distribution. If reviewers use many different words, entropy would be high. If a reviewer reuses a smaller

[3]http://www.yelp.com/dataset_challenge

subset of words, the entropy of word distribution would be low, as this would represent a less uniform distribution over word types.

A third measure is the average information encoded in the reviewer's word use, which we'll call average unigram information (*AUI*). Information, as described above, is a measure of the number of bits a word encodes given its frequency in the overall corpus. Reviews with higher information use less frequent words, thus offering more specific and less common verbiage in describing a business.

A fourth measure is one more commonly used in studies of informational structure of language, which we'll call the average conditional information (*ACI*). This is a bit-based measure of a word based on its probability conditioned on the prior word in the text. In other words, it is a measure of the bits encoded in a given bigram of the text. We compute the average bits across bigrams of a review, which reflect the uniqueness in word combinations.[4]

Finally, we extract two crude measures of information variability by calculating the standard deviation over AUI and *ACI*, which we call unigram informational variability (*UIV*) and conditional informational variability (*CIV*) respectively. Both measures are a reflection of how stable the distribution is over a reviewer's average unigram and bigram bit values. These measures relate directly to uniform information density (Jaeger, 2010; Jaeger & Levy, 2007). A very uniform distribution of information is represented by a stable mean and lower *UIV*/*CIV*; a review with unigram or bigram combinations that span a wide range of informativeness induces a wider range of bit values, and thus a higher *UIV*/*CIV* (less uniform density). A summary of these measures appears in Table 3.1. Punctuation, stop words and spacing were removed using the `tm` package in `R` before information-theoretic measures were obtained.[5]

3.3.3 Social Networks

One benefit of the big-data approach we take in this chapter is that we can pose our questions about language and social structure using the targeted unit of analysis of social networks themselves. In other words, we can sample networks from the Yelp dataset directly, with each network exhibiting network scores that reflect a certain aspect of local social structure. We can then explore relationships between the information-theoretic measures and these social-network scores.

We sampled 962 unique social networks from the Yelp dataset, which amounted to approximately 38,000 unique users and 450,000 unique reviews. Users represent nodes in social networks and were selected using a selection and connection algorithm also shown in 3.1. We start by choosing a random user who has between 11 and 20 friends in the data set (we chose this range to obtain networks which were not too small or too large as to be computationally cumbersome). After we chose that user, we connected all her friends and then expanded the social network by one degree; randomly selecting 10 of the her friends and connecting up to 10 of her friend's friends to the network. We then interconnected all users in this set (shown as the first-degree nodes and connections in 3.1). We conducted this same process of finding friends of these first-degree nodes, and then interconnected those new nodes of the second degree. In order to make sure networks did not become too large, we randomly sampled up to 10 friends of each node only. 50% of all networks fell between 89-108 reviewers in size, and the resulting nets reveal a relatively normal distribution of network metrics described in the next section.

3.3.4 Network Measures

A variety of different network measures were used to quantify the structure of each network. We consider two different categories of network structures: simple and complex. A

[4]Previous research calls this Information Density and uses this as a measure of Uniform Information Density. We use the name Average Conditional Information given the breadth of information-theoretic measures used in this study.

[5]Note: *AUI* and *ACI* were calculated by taking only the unique *n*-grams.

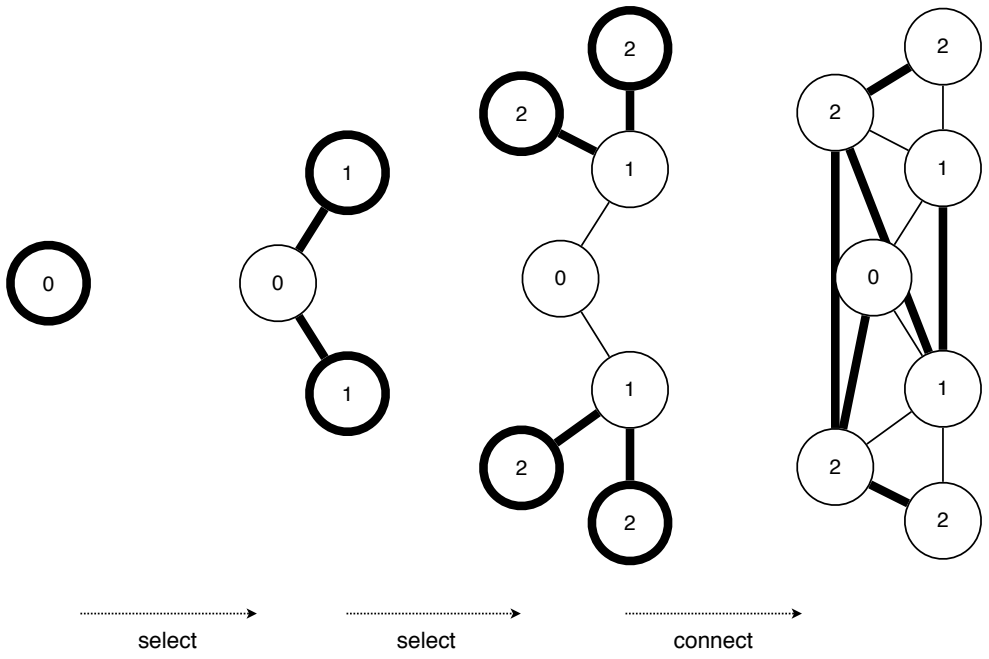


Figure 3.1: A simplified visualization of the network building process

”0”-degree was chosen at random among those with between 11 and 20 friends. We then connected these individuals. Then, from 10 randomly chosen friends of the seed node, we chose up to 10 friends of friends and connected them. Following this, we interconnected the whole set.

summary of all seven (2 simple, 5 complex) network measures appears in Table 3.2. We used two simple network measures: The number of reviewers in a network, or *nodes*, and the number of friendship connections between reviewers, or *edges*.

We considered five complex network measures. The first measure, network *degree*, is the ratio of edges to nodes. This provides a measure of connectivity across a network. High-degree networks have a higher edge-to-node ratio than lower degree networks.

The second measure, network *transitivity*, determines the probability that two adjacent nodes are themselves connected (sometimes termed the "clustering coefficient"). Groups of three nodes, or triples, can either be closed (e.g., fully connected) or open (e.g., two of the three nodes are not connected). The ratio of closed triples to total triples provides a measure of the probability that adjacent nodes are themselves connected. High transitivity occurs when the ratio of closed-to-open triples is close to one.

A third measure, network *betweenness*, determines the average number of shortest paths in a network that pass through some other node. The shortest path of two nodes is the one that connects both nodes with the fewest edges. A pair of nodes can have many shortest paths. A node's betweenness value is the sum of the ratio of a pair of node's shortest paths that pass through the node, over the total number of shortest paths in the network. We determined network betweenness by taking the average node betweenness for all nodes in the network. Effectively, this provides a measure of network efficiency. The higher a network's betweenness, the faster some bit of new information can travel throughout the network.

A fourth measure stems from node *centrality* which determines the number of connections a single node has with other nodes. Centrality can also be determined for the whole network, known as graph centrality. Graph centrality is the ratio of the sum of the absolute value of the centrality of each node, over the maximum possible centrality of each node (L. C. Freeman, 1978). Node centrality is greatest when a single node is connected to all other nodes, whereas graph centrality is greatest when all nodes are connected to all other nodes. Information is thought to travel faster in high-centrality networks. Here we use graph centrality only. From this point on we will refer to to graph centrality simply as centrality. Network betweenness and network centrality share common theoretical assumptions, but quantify different structural properties of a network.

Our fifth and final measure determines whether the connections between nodes in a network share connectivity at both local and global scales. Scale free networks display connectivity at all scales, local and global, simultaneously (Dodds, Watts, & Sabel, 2003). A network is said to be *scale free* when its degree distribution (i.e., the number of edge connections per each node) fits a power law distribution. Networks that are less scale free are typically dominated by either a local (a tightly connected set of nodes) or global connectivity (randomly connected nodes). Networks that exemplify differences in complex structures are presented in Figure 3.2.

3.3.5 Additional Measures

Individual reviews were not quantified independently. Instead, all reviews from a single individual were concatenated into one document. This allowed for information-theoretic measures to be performed over a single user's total set of reviews. The average information of a network was then computed by taking the average across all individuals (nodes) in the network. Such an analysis affords testing how the structure of an individual's social network impacts that individual's overall language use. However, due to the nature of how our information-theoretic measures were determined, individuals who wrote well over one hundred reviews were treated the same as those who wrote merely one. This introduces a possible bias since information measures are typically underestimated when using non-infinite sample sizes (as in the case of our information measures). While we control for certain measures such as the average reviewer's total review length and network size, additional biases may occur due to the nature of how each measure was determined (e.g., averaging across reviewers with unequal length reviews). To

Table 3.2: Summary of the measures quantifying a network's structure

Measure	Definition	Description
Nodes	$Nodes$	Number of individuals in the network
Edges	$Edges$	Number of node to node connections; vertices in a network
Degrees	$\frac{Edges}{Nodes}$	The ratio of connections to nodes in a network
Transivity	$\frac{N\ closed\ Triples}{N triples}$	The average number of completely connected triples given the total number of triples in a network.
Betweeness	$\frac{\Sigma\left\{\sum_{s\neq t\neq v}\frac{SP_{st}(V)}{SP_{st}}\right\}}{N}$	SP_{st} is the number of total shortest paths from node s to node t. $SP_{st}(V)$ is the number of shortest paths from s to t that pass through node V. The sum for all shortest paths for all nodes determines the betweenness of node V. We take the average betweenness of each node for all nodes N in a network.
Centrality	$C_x = \frac{\sum_{i=1}^{N}\lvert C_x(p^*)-C_x(p_i)\rvert}{Max\sum_{i=1}^{N}\lvert C_x(p^*)-C_x(p_i)\rvert}$	C_x is the graph level centrality defined as the sum of the absolute difference between the observed maximum central node $C_x(n^*)$ and all other node centrality measures $C_x(n_i)$ over the theoretical maximum centrality of a network with the same number of nodes. Since, this is a measure of the maximum possible centrality and actual centrality, graph level centrality will always fall between 0 (low centrality) and 1 (high centrality).
Scale Free	$f(x) = x^{-\alpha}$	α is the exponent characterizing the power law fit predicted by the degree distribution x. α is always greater than 1 and typically falls within the range of 2 ¡ α ¡ 3, but not always.

Note: N = number of words in a review; $p(w)$ = probability of word w; w_i = ith word of a review; UI_j = set of unigram information scores for each word of a given review; CI_j = set of conditional information scores for each word of a given review.

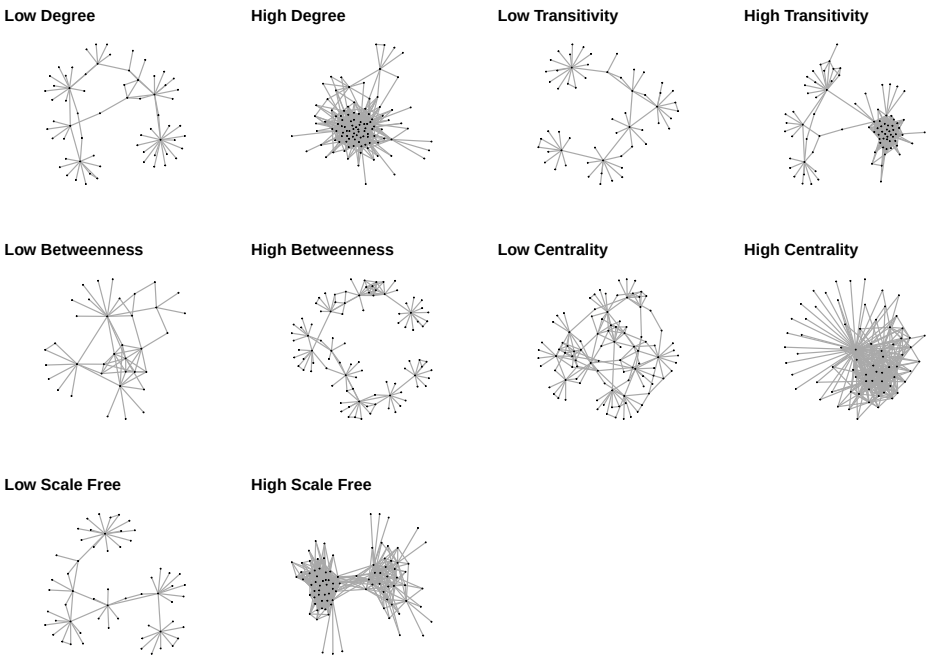


Figure 3.2: Example Yelp networks with high/low structural properties

address these concerns two additional measures, *(1)* a gini coefficient and *(2)* a random review baseline— to assess the reliability of our analyses—were used. They are described below.

Gini coefficient. The Gini coefficient ($range = [0,1]$) was originally developed to assess the distribution of wealth across a nation:

$$G = \frac{n+1}{n} - \frac{2\sum_1^n (n+1)x_i}{n\sum_1^n x_i} \tag{3.2}$$

As the coefficient approaches zero, wealth is thought to approach greater equality. As the coefficient approaches one, more of the nation's wealth is thought to be shared amongst only a handful of its residents. We use the Gini coefficient to assess the distribution of reviews across a network. Since each node's reviews were concatenated, given only one value for each information-theoretic measure, certain reviewer's measures will be more representative of the linguistic distributions. The Gini coefficient provides an additional test as to whether a network's average information is influenced by the network's distribution of reviews.

Random review baseline. A random network baseline provides a baseline value to compare coefficient values from true networks. Baseline information measures were computed by randomly sampling (without replacement) the same number of reviews written by each reviewer. For example, if a reviewer writes five reviews, then five random reviews are selected to take their place. These five reviews are then deleted from the pool of total reviews used throughout all networks. This ensures that the exact same reviews are used in baseline reviews. We did not go so far as to scramble tokens within reviews. While this would provide a sufficient baseline, obtaining the true information-theoretic measures of each review, without token substitution, provides a more conservative measure. The random review baseline was used to compare all significant true network effects as an additional measure of reliability.

3.4 Predictions

This chapter presents a broad thesis, and we test it using the Yelp data: Social-network structure will relate in interesting ways to patterns of language use. Though this is a broad expectation, it is not a specific prediction, and we do not wish to take a strong stance on specific hypotheses here, giving the reader the impression that we had conceived, in advance, the diverse pattern of results that we present below. Instead, big data is providing rich territory for new exploration (Lohr, 2012). The benefit of a big-data approach is to identify interesting and promising new patterns or relationships and, we hope, encourage further exploration. As we show below, even after controlling for a variety of collinearities among these measures, the broad thesis holds. Regression models strongly relate social-network and information-theoretic measures. Some of the models show proportion variance accounted for at above 50%. Despite this broad exploratory strategy, a number of potential predictions naturally pop out of existing discussion of information transmission in networks. We consider three of these here before moving into the results.

One possibility is that more scale-free networks enhance information transmission, and thus would concomitantly increase channel capacity of nodes in the network. One might suppose that in the Yelp context, a more local scale-free structure might have a similar effect: An efficient spread of information may be indicated by a wide diversity of information measures, as expressed in the UIV and CIV measures.

A second prediction—not mutually exclusive from the first—draws from the work on conceptual entrainment and imitation in psycholinguistics, is that densely connected nets may induce the least AUI. If nodes in a tightly connected network infect each other with common vocabulary, then this would reduce the local information content of these words, rendering them less unique and thus show the lowest entropy, AUI, and so on. One may expect something similar for transitivity, which would reflect the intensity of local mutual interconnections (closed triples).

However, the reverse is also possible. If language users are more densely connected it may be more likely that they have established a richer common ground overall. If so, language use may contain more information-dense words specific to a shared context (Doyle & Frank, 2015). A fourth prediction is that more network connectivity over a smaller group (higher-network degree) may afford more complex language use, and so lead to higher AUI and ACI.

A final prediction comes from the use of information theory to measure the rate of information transmission. When a speaker's message is more information-dense, it is more likely that it will also be more uniform. Previous findings show speakers increase their speech rate when presenting low information-dense messages, but slow their speech rate for information-dense messages (Pellegrino, Coupé, & Marsico, 2011). It may be that any social structure that leads to increases in information density simultaneously decreases information variability.

3.5 Results

3.5.1 Simple measures

The confidence intervals (99.9%CI) of five multiple regression models, where nodes, edges and the Gini coefficient were used to predict each information-theoretic measure, are presented in Table 3.3. We set a conservative criterion for significance ($p < .001$) for all analyses. Only those analyses that were significant are presented. Crucially, all significant effects of independent variables were individually compared to their effects on the random review baseline. To do this, we treated the random review baseline and true network reviews as two levels of the same variable: "true_baseline". Using linear regression we added an interaction term between independent network variables and the true_baseline variable. A significant interaction is

Table 3.3: Lexical measures as predicted by Nodes, Edges and Gini Coefficient

	Nodes	*Edges*	*Gini Coef*	*F-statistic*
Length	*n.s.*	(.09, .27)	(.30 , .43)	$F(3,958) = 125$ $R^2 = .28$, $R^2_{adj} = .28$
$RI - Ent_{residual}$	*n.s.*	(.10, .15)τ	(−.16, −.12)τ	$F(3,958) = 446.6$ $R^2 = .58$, $R^2_{adj} = .58$
$AUI_{residual}$	(−.05,−.01)τ	(.10, .12)τ	(−.12, −.09)τ	$F(3,958) = 391$ $R^2 = .55$, $R^2_{adj} = .55$
$ACI_{residual}$	*n.s.*	(.07, .12)τ	(−.04, −.01)	$F(3,958) = 154.1$ $R^2 = .33$, $R^2_{adj} = .32$
$UIV_{residual}$	(−.01,−.001)τ	(.001, .01)τ	(.004, .01)	$F(3,958) = 39.6$ $R^2 = .11$, $R^2_{adj} = .11$
$CIV_{residual}$	*n.s.*	(−.003, 0)	(.002, .004)	$F(3,958) = 27.99$ $R^2 = .08$, $R^2_{adj} = .08$

Note: Only the Mean and 99.9% Confidence Intervals for each IV with p¡.001 are presented. The τ *symbol denotes all network effects that were significantly different from baseline network effects (*$p < .001$*).Multiple linear regressions were performed in* `R:` `lm(DV Nodes+Edges+Gini)`

demarcated by "τ" in Table 3.3 and 3.4. The effects of these network variables on information-theoretic measures are significantly different in true networks compared to baseline networks. This helps to ensure that our findings are not simply an artifact of our methodology.

All variables were standardized (scaled and shifted to have $M = 0$ and $SD = 1$). Additionally, the number of words (*Length*) was log transformed due to a heavy-tailed distribution. All other variables were normally distributed. Because length correlates with all information-theoretic measures and *UIV* and *CIV* correlate with the mean of *AUI* and *ACI*, respectively (due to the presence of a true zero), the mean of each information measure was first predicted by *Length* while *UIV* and *CIV* were also predicted by *AUI* and *ACI*. The residual variability of these linear regression models was then predicted by nodes, edges and the *Ginicoefficient*. The purpose of residualization is to further ensure that observed interactions are not due to trivial collinearity between simpler variables (*Length*, *Nodes*) and ones that may be subtler and more interesting (*CIV*, *centrality*, etc.).[6]

The number of nodes provides a crude measure of network size, edges, network density and the Gini coefficient (the distribution of reviews across the network). Importantly, no correlation exists between the Gini coefficient with either edges or nodes. And, although a strong correlation exists between nodes and edges ($r = .67$, $t(960) = 28.23$, $p < .0001$), in only two instances, *AUI* and *UIV*, did nodes account for some portion of variance. As nodes increased, average unigram information, along with average unigram variability, decreased. However, only the relationship between nodes and average unigram information was significantly different between the true network and the random review baseline. In all cases, save conditional information variability (*CIV*), a significant proportion of variance in information measures was accounted for by edges, and in all but length and *CIV*, the relationship between information measures and edges was significantly different between the true network and the random review baseline (3.3a presents an example interaction plot between *ACI*, edges and true_baseline

[6]Our approach toward controlling for collinearity by residualizing variables follows that of previous work (Jaeger, 2010). However, it is important to note the process of residualizing to control for collinearity is currently in debate (Wurm & Fisicaro, 2014). It is our understanding that the current stringent use of this method is warranted provided it stands as a first pass toward understanding how language use is influenced by network structures.

Table 3.4: Information-Theoretic measures predicted by complex network measures

	Degree	*Transitivity*	*Betweeness*	*Centrality*	*Scale Free*	*F-statistic*
Length	*n.s.*	(.20, .42)	(−.28 , −.08)	(.01, .16)	(*n.s.*)	$F(5,956) = 33.3$ $R^2 = .15, R^2_{adj} = .14$
$RI - Ent_{residual}$	(.04, .13)τ	(−.11, −.03)	*n.s.*	*n.s.*	*n.s.*	$F(5,956) = 21.5$ $R^2 = .10, R^2_{adj} = .10$
$AUI_{residual}$	(.02,.09)τ	*n.s.*	*n.s.*	(.004, .05)τ	*n.s.*	$F(5,956) = 10.4$ $R^2 = .05, R^2_{adj} = .05$
$ACI_{residual}$	(.01, .07)τ	(−.07, −.01)τ	*n.s.*	*n.s.*	*n.s.*	$F(5,956) = 11.6$ $R^2 = .06, R^2_{adj} = .05$
$UIV_{residual}$	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*
$CIV_{residual}$	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*

Only the Mean and 99.9% Confidence Intervals for each IV with $p < .001$ are presented. All reported values were significant ($p < .001$). The τ symbol denotes all network effects significantly different from baseline network effects.

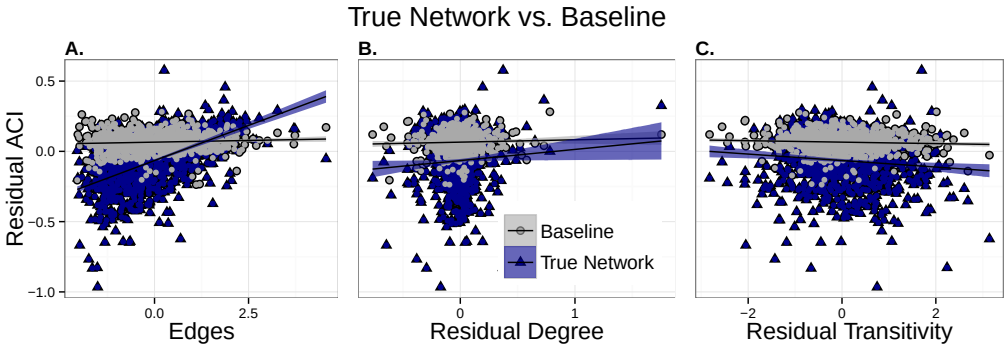


Figure 3.3: Network Measures for True and Baseline networks by Across Conditional Information Density

All four plots show significant interactions for variables in True networks compared to baseline networks. Linear regression models with interaction terms were used in `R:lm(DV IV+true_baseline+IV*true_baseline)`

measures). Finally, the Gini coefficient accounted for a significant portion of variance for all information measures, but only for $RI - Ent$ and AUI did it have a significantly different relationship between the true network and the random review baseline. One explanation may be that more unique language use may naturally occur when more individuals contribute more evenly to the conversation. Another possibility is that networks with less even review distributions are more likely to have more reviews, suggesting that a larger number of reviewer's language use is more representative of the linguistic distribution of reviews. A simple linear regression analysis reveals the Gini coefficient accounts for a significant portion of variance in the total number of reviews in each network ($R^2_{adj} = .21$, $F[1,960] = 249.6$, $p < .001$, 99.9% CI [.25, .38]), increasing as the number of reviews increase.

We interpret these results cautiously considering it is a first step toward understanding what aspects of a network relate to language use. The results suggest changes in population size and connectivity occur alongside changes in the structure of language use. Speculatively, the individual language user may be influenced by the size and connectivity of her network. When the size of her network increases, the words she uses may be more frequent. However when connectivity increases, the words she uses may be of low frequency, and therefore more information dense. This supports current work that shows how a shared common ground may lead to an increase in information dense word use (Doyle & Frank, 2015). This is further explored in the discussion.

Though we find significant effects, how network size and connectivity influence information density and channel capacity, and how different ways of interpreting information (as we have done here) interact with simple network measures is unclear. Generally, these results suggest that word choice may relate to social-network parameters. overall

3.5.2 Complex measures

The complex network measures, *centrality*, *degree* and *scale free*, were log transformed for all analyses due to heavy-tailed distributions. Given the larger number of network variables and their use of similar network properties such as number of nodes or edges, it is possible that some complex network measures will be correlated. To avoid any variance inflation that may occur while using multiple predictors, we determined what variables were collinear using a variance inflation factor (`VIF`) function in `R`. We first used nodes and edges to predict the variance of each complex network measure. We factor this out by taking the residual of

each model and then used the VIF function from the R library car to determine what complex network measures exhibited collinearity. Using a conservative VIF threshold of 5 or less used in our model was at risk of any collinearity that would have seriously inflated the variance. All VIF scores were under the conservative threshold for all complex network variables and are therefore not reported. Residuals of complex network measures, having factored out any variance accounted for by nodes and edges, were used to predict each information-theoretic measure presented in Table 3.4.[7]

One or more complex measures accounted for a significant proportion of variance in each information density measure. Certain trends are readily observed across these models. Specifically, word length increased as network transitivity and centrality increased and decreased as network betweenness increased, however no network measure effects were significantly different from random review baseline effects (significance marked by the τ symbol in Table 3.4). Additionally, $RI-Ent$ and AUI and ACI increased as network degree increased accounting for between $5-10\%$ of the variance in each measure. The relationship between network degree and corresponding information measures in true networks was significantly different from baseline. This was also the case for network centrality for both AUI and network transitivity for ACI. 3.3 presents interaction plots for residual ACI by degree (3.3b) and residual transitivity (3.3c) between true and random review baseline networks. Complex network measures did not share a significant relationship with UIV or CIV.

It is clear that certain network structures predict differences in information density measures even after stringent controls were applied to both information and network measures. Specifically, support for higher information dense messages may be the result of networks that exhibit high global connectivity driven by increases in specific network properties, namely, network degree and centrality. This further supports previous work showing that a shared common ground may bring about higher information-dense language use. Specifically, networks that exhibit a more centralized nucleus and are more densely connected (higher degree) may be more likely to share a similar common ground amongst many members of the group. If so, a shared common ground may result in more unique language use. Yet, networks that exhibit close, niche-like groupings exemplified by high network transitivity, may infect its members with the same vocabulary, decreasing the overall variability in language use. Further analysis is necessary to unpack the relationship that different social-network structures have with language use.

3.5.3 Discussion

We find pervasive relationships between language patterns, as expressed in information-theoretic measures of review content, and social network variables, even after taking care to control for collinearity. The main findings are listed here:

- Reviewers used more information-dense words ($RI-Ent$, AUI) and bigrams (ACI) in networks with more friendship connections.
- Reviewers used more information-dense words ($RI-Ent$, AUI) in networks that have a lower Gini coefficient; networks where reviews were more evenly distributed.
- Reviewers use more information-dense words ($RI-Ent$, AUI) and bigrams (ACI) as network degree (ratio of friendships connections to number of individuals in the network) increased and as individuals in the network were grouped more around a single center (AUI only).

[7] The variance inflation acceptable for a given model is thought to be somewhere between 5 and 10 (Craney & Surles, 2002). After the variance predicted by nodes and edges was removed from our analyses, no complex network measure reached the variance inflation threshold of 5.

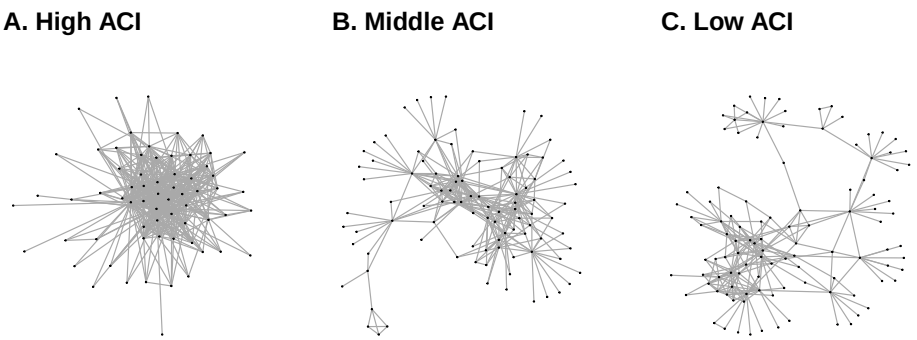


Figure 3.4: Yelp networks at the tails of complex network measure distributions exhibiting High (*A*) Middle (*B*) and Low (*C*) Average Conditional Information (*ACI*)

- Reviewers used less information-dense bigrams as the number of local friendship connections increased (e.g., network transitivity).
- Unigram information variability (*UIV*) was higher with higher connectivity; channel capacity was less uniform in networks with more friendship connections.

The predictions laid out at the end of the Methods section are somewhat borne out. Scale-free networks do not appear to have a strong relationship among information-theoretic scores, but networks that exhibit higher transitivity do lead to lower information-dense bigrams (though not significant for any other information measure) and, while more connections lead to higher information density, they do not lead to lower information variability. Indeed, when considering the last finding, the opposite was true: Networks with higher connectivity used more information-dense words at a more varied rate. Although this was not what we predicted, it is in line with previous work supporting the notion that certain contextual influences afford greater resilience to a varied rate of information transmission (Vinson & Dale, 2014b). In this case, more friendship connections may allow for richer information-dense messages to be successfully communicated less uniformly.

We found support for two predictions; *(1)* high transitivity networks lead to less information-dense bigram language use and *(2)* high degree networks tend to exhibit higher information density. In addition, more centralized networks also lead to higher information-dense unigram language use. The first prediction suggests that networks where more local mutual interconnections exist may be more likely to infect other members with similar vocabulary. That is, more local connectivity may lead to more linguistic imitation or entrainment. Here we merely find that the structure of reviewers' language use is similar to one another. It is possible that similarities in linguistic structure reveal similarities in semantic content across connected language users, but future research is needed to support this claim.

Support for the second prediction suggests that users adapt their information-dense messages when they are more highly connected. This effect can be explained if we assume that certain social network structures afford groups the ability to establish an overall richer common ground. Previous work shows that increased shared knowledge leads to more information dense messages (Doyle & Frank, 2015; Qian & Jaeger, 2012). It may be that increases in network degree and centrality enhance network members' abilities to establish a richer common ground, leading to more information dense messages. One possibility may be that certain networks tend to review similar types of restaurants. Again, further exploration into how the number of

friendship connections, network degree and centrality impact information density and variability is needed to determine the importance of specific network properties in language use. Figure 3.4(a-c) provide example networks that exhibit low, middle and high network ACI given the specific network structures that predict ACI above (e.g., increases in network degree and decreases in transitivity).

3.6 General Discussion

In this chapter we explored how language use might relate to social structure. We built 962 social networks from over 200,000 individuals who collectively wrote over one million online customer business reviews. This massive, structured dataset allowed testing how language use might adapt to structural differences in social networks. Utilizing big data in this way affords assessing how differences in one's local social environment might relate to language use and communication more generally. Our findings suggest that as the connectivity of a population increases, speakers use words that are less common. Complex network variables such as the edge-to-node ratio, network centrality, and local connectivity (e.g., transitivity) also predict changes in the properties of words used. The variability of word use was also affected by simple network structures in interesting ways. As a first exploration our findings suggest local social interactions may contribute in interesting ways language use. A key strength of using a big-data approach is in uncovering new ways to test theoretical questions about cognitive science, and science in general. Below we discuss how our results fit into the broader theoretical framework of understanding what shapes language.

When controlling for nodes, edges and review length, many R^2 values in our regression models were lowered. However, finding that some variability in language use is accounted for by population connectivity suggests language use may be partly a function of the interactions among individuals. Both network degree, centrality and transitivity varied in predictable ways with information measures. Mainly, as the number of connections between nodes increased and as the network became more centralized the use of less frequent unigrams (AUI) increased. Interestingly, networks that exhibit high connectivity and greater centrality may have more long-range connections. A growing number long-range connections may lead to the inclusion of individuals that would normally be farther away from the center of the network. Individuals in a network with these structural properties may be communicating more collectively; having more readily established a richer common ground. If so, higher information density is more probable, as the communication that is taking place can become less generic and more complex. Additionally, networks with higher local connectivity, or high transitivity tend to use more common language, specifically bigrams. This again may be seen as supporting a theory of common ground, that individuals with more local connectivity are more likely to communicate using similar terminology, in this case, bigrams. Using a big-data approach it is possible to further explore other structural aspects of one's network that might influence language use.

While we merely speculate about potential conclusions, it is possible to obtain rough measures of the likelihood of including more individuals at longer ranges. Specifically, a network's diameter —the longest stretch of space between two individual nodes in any network —may serve as a measure of the distance that a network occupies in socio-cultural space. This may be taken as a measure of how many strangers are in a network, with longer diameters being commensurate with the inclusion of more strangers.

It may be fruitful to explore the impact of a single individual on a network's language use. We do not yet explore processes at the individual level, opting instead to sample networks and explore their aggregate linguistic tendencies. Understanding the specifics of individual interaction may be crucial toward understanding how and why languages adapt. We took an exploratory approach and found general support for the idea that network structure influences certain aspects of language use, but we did not look for phonological or syntactic patterns; in

fact our analysis could be regarded as a relatively preliminary initial lexical distribution analysis. However, information finds fruitful application in quantifying massive text-based datasets and has been touted as foundational in an emerging understanding of language as an adaptive and efficient communicative system (Jaeger, 2010; del Prado Martın, Kostić, & Baayen, 2004). In addition, previous work investigating the role of individual differences in structuring one's network are important to consider. For instance, differences in personality, such as being extroverted or introverted, are related to specific network-level differences (Kalish & Robins, 2006). It is open to further exploration as to how information flows take place in networks, such as through hubs and other social processes. Perhaps tracing the origin of the network by determining the oldest reviews of the network and comparing these to the network's average age may provide insight into the importance of how certain individuals or personalities contribute to the network's current language use.

We see the current results as suggestive of an approach toward language as an adaptive and complex system (Beckner et al., 2009; Lupyan & Dale, 2015). Our findings stand alongside previous research that reveals some aspect of the structure of language adapts to changes in one's socio-cultural context (Klingenstein et al., 2014; Kramer, Guillory, & Hancock, 2014; Lupyan & Dale, 2010; Vilhena et al., 2014). Since evolution can be thought of as the aggregation of smaller adaptive changes taking place from one generation to the next, finding differences in language within social networks suggests languages are adaptive, more in line with shifts in social and cultural structure than genetic change (Chater et al., 2009; Gray & Atkinson, 2003). The results of this study suggest that general language adaptation may occur over shorter time scales, in specific social contexts, that could be detected in accessible big-data repositories (Stoll, Zakharko, Moran, Schikowski, & Bickel, 2015). The space of communicable ideas may be more dynamic, adapting to both local and global constraints at multiple scales of time. A deeper understanding of why language use changes may help elucidate what ideas can be communicated when and why. The application of sampling local social networks provides one method toward understanding what properties of a population of speakers may relate to language change over time —at the very least, as shown here, in terms of general usage patterns.

Testing how real network structures influence language use is not possible without large amounts of data. The network sampling technique used here allows smaller networks to be sampled within a much larger social-network structure. The use of big data in this way provides an opportunity to measure subtle and intricate features whose impacts may go unnoticed in smaller-scale experimental datasets. Still, we would of course recommend interpreting initial results cautiously. The use of big data can provide further insight into the cognitive factors contributing to behavior, but can only rarely be used to test for causation. To this point, one major role the use of big data plays in cognitive science, and one we emphasize here, is its ability to provide a sense of direction and a series of new hypotheses.

Chapter 4

Efficient n-gram Analysis with `cmscu`

4.1 Introduction

"Big data" collection and analysis are now at the forefront of modern science and business, with daily data collection equal to that of 90% of all data collected in the past two years (McAfee et al., 2012), which comprises over 2.7 zettabytes (10^{21} bits). Keeping pace with the scale and speed of modern data collection necessitates the development of computational tools capable of efficiently analyzing larger and larger data sets. Hardware has rapidly evolved to enable such large-scale computations; for example, the time to assemble an entire human genome, just under one week, once took two years (Sagiroglu & Sinanc, 2013). To maximize the utilization of modern hardware, however, calculations must typically be expressed in "low-level" languages such as Fortran or C++, more or less directly exposing the hardware to the programmer at the cost of code simplicity and clarity. More expressive scripting languages such as R and Python allow scientists to more directly express their calculations and theories in a hardware-agnostic way, but at the (sometimes significant) expense of code runtime and not being able to immediately leverage the latest hardware advances.

Because of this, tools that facilitate analysis of large data sets could greatly accelerate research in behavioral science. Many (if not most) behavioral scientists are trained only in scripting languages and the relevant statistical packages. Thus the analysis of larger data sets falls to computer scientists and engineers who often lack the background and training in the behavioral sciences. Further integrating these fields through big data and analysis tools has much promise within both research and the applied domains. For example, data sets continue to be released to the public with companies benefiting from 'dataset challenges' that crowd source solutions to computational problems. Netflix famously created a dataset challenge paying out one million dollars to any team who could improve their current recommendation system by 10%. Yelp Inc. continues to release more and more reviews from their database, and pays out $5,000 to students who simply use the data in interesting ways. In fact, there are entire websites dedicated to advertising dataset challenges (e.g., Kaggle.com). The problem is that behavioral science training rarely includes efficient programming techniques to harness the raw power of these larger data sets, often sacrificing computational efficiency for ease of programming. In order to obtain insights from larger data sets, behavioral scientists —with their important domain-specific knowledge —must be able to engage with ever larger data sets in a meaningful way.

The scripting language `R` is the preferred computational and statistical language of many behavioral scientists, having rich documentation and an accessible programming environment. However, it is not very performant relative to other solutions (Simmering, 2013).

Nonetheless, the use of high-level scripting languages, and in particular `R`, has been encouraged for many current and past research agendas. Historically, this has not been a problem as the size of one's data set and complexity of analyses have typically been guided by highly controlled experimental paradigms. Such data sets require very little computational power to uncover phenomena from few participants. However, the increasing size and number of freely available data sets, as well as the desire to provide ecologically valid results (Roy, Frank, DeCamp, Miller, & Roy, 2015; Lazer et al., 2009), challenges the efficacy of this trade-off. Indeed, being capable of harnessing large data sets will also help to accommodate the current needs of the behavioral sciences, such as openness and replicability (Nosek, Spies, & Motyl, 2012; Schmidt, 2009; Pashler & Wagenmakers, 2012; Zwaan & Pecher, 2012), without loss of scientific productivity (Ramscar, Shaoul, Baayen, & Tbingen, 2015). This new demand requires new tools that can be easily implemented by behavioral scientists trained to uncover interesting behavioral phenomena.

In this paper, we introduce the application of a prominent data-reduction technique for handling massive amounts of text. This technique permits efficient approximation of text information from a very large corpus. The `R` library we introduce, `cmscu`, could thus open new avenues of analysis for behavioral scientists, but the library and its application also recommend some broad methodological lessons. We aim to elucidate three key methodological observations. First, in Section 2, R-package `cmscu` 4.2, we describe how so-called "sketch" techniques help process large amounts of data while efficiently using memory resources and argue these are critical for the coming 'big data' age in the behavioral sciences (Griffiths, 2015). Next, in Section 3, Information-Theoretic Structure of Yelp Reviews 4.3, we demonstrate a fruitful domain in which such a strategy can apply —the Information-Theoretic analysis of language structure in corpora. As an example, due to the efficiency now available with `cmscu`, we quantify the structure of a language using the sophisticated *Modified Kneser-Ney smoothing* algorithm (Kneser & Ney, 1995). Finally, using a dataset from Yelp, Inc., we show that our library and its use in implementing sophisticated Information-Theoretic algorithms permit wide ranging exploration of the statistical properties of language use and its relationship other behavioral phenomena. In the General Discussion, we revisit important messages about cross-disciplinary interactions and suggest that adopting a wide range of tools from various disciplines can help to ensure that behavioral scientists find efficient solutions for their problems, while giving computational scientists and engineers exciting behavioral problems in which to apply their research.

Here, we briefly motivate why adopting more efficient techniques is crucial to maintaining the current pace of progress in the behavioral sciences. To do this we demonstrate its application in one of the most common quantitative models of language: n-grams.

***n*-gram models of language** The study of language is a key topic in the psychological sciences as noted in the introduction. To review, n-gram models of language use have been used to quantify the statistical structure of its use and are shown to influence and be influenced by many psychological factors. As explained in the Introduction, Information Theory posits that a word carries some *amount* of information, measured in *bits*, proportional to its $-\log_2$ probability of occurrence given some "context" (e.g., a corpus of words):

$$I(w_i) = -\log_2 p(w_i|w_1, w_2, ...w_{i-1}). \tag{4.1}$$

The number of bits in word $I(w_i)$ is dependent on the frequency of its occurrence *after* some other word(s), or its maximum likelihood estimate. In the behavioral and computational sciences, this measure and related measures are very useful for exploring language production and processing. However, computing reliable estimates of these measures requires the analysis of massive amounts of text. In addition, it may be very useful to compute these measures over ad hoc data sets of psychological relevance (e.g., education, business, etc.). Doing so requires flexible tools that allow behavioral researchers to estimate such measures in their existing com-

putational environment. This would greatly enhance the availability of such Information Theory concepts.

Currently, programs and methods used by psychologists, specifically the `tm` package in `R`, do not scale well with increasingly large data sets or longer n-grams, thus leaving a powerful tool out of reach. Below we detail the development and use of a new n-gram package, `cmscu`, developed in part to analyze a large Yelp, Inc. dataset previously intractable with the standard `tm` package and its `DocumentTermMatrix` object in `R`. Following its description, we provide the results of an n-gram analysis motivated by current and ongoing research surrounding Information Theory.

4.2 R-package `cmscu`

The analysis of n-grams requires two fundamental operations: store and query. Given an n-gram ω, we must be able to store and update the count associated with ω, $store(\omega)$, and query the same count, $query(\omega)$. There are many possible ways to implement such functionality; however, the large nature of our intended dataset limits the feasibility of most standard approaches. In order to be fast, the data structure (or at least its "active parts") should be able to reside entirely within local memory (RAM). This means that we cannot fully store the actual n-grams themselves, but rather a compressed representation of them.

The need for compression suggests a hash table implementation, "compressing" strings by representing them as a single integer *computed* from the string itself (Cormen, 2009). For example, we may use a look-up table for a trigram such as `the_cannibal_consumes` by referencing a single integer —an index in an array —thus saving significant space in the frequency table. However, such hash tables store the string itself along with its associated data in order to resolve possible hash collisions, where two distinct strings map to the same integer —a necessary consequence of the mathematical pigeon-hole principle. As the size of the corpus increases, the space required to store all strings will exceed the size of RAM thus requiring swapping memory from the hard disk, which is often the single largest slowdown in large programs.

By relaxing the "correctness" of our stored value (in a controlled way) a host of fast and efficient algorithms become available. So-called streaming or sketch algorithms rely on *probabilistic* guarantees of accuracy: in this framework, $query(\omega)$ will return a value according to a confidence interval rather than the fixed, precise number. We chose the Count-Min-Sketch with conservative update procedure (`CMS-CU`) (Cormode & Muthukrishnan, 2005) for its simplicity in implementation and its proven effectiveness (Goyal et al., 2012). This approach has wide applications, and approximate or "sketch" algorithms in general have become quite common in the realm of computational and data science for summarizing and performing other computations on massive and real-time data (Cormode & Muthukrishnan, 2011), the most common example being the conceptually-related Bloom filter (Song, Dharmapurikar, Turner, & Lockwood, 2005). By treating a corpus as a stream of text data, an approximate algorithm such as Count-Min-Sketch can provide a estimate of n-gram frequencies up to some desired accuracy depending on core parameters in the algorithm —namely the size of the table that will provide the information for the estimation of frequencies. This allows a researcher to *choose a balance* between probabilistic guarantees and available computational resources, effectively trading off statistical power for the ability to study larger datasets *on the same hardware*.

The CMS-CU sketch algorithm works in the following way: rather than representing some n-gram in a traditional hash table, containing an entry for each distinct n-gram, the count data of the n-grams are tabulated in the stream of text using a w-by-d table ("width and depth"). Every n-gram receives an entry on each row of this table, and the particular entry in each row is determined by a statistically independent hash function. Storing an n-gram consists of incrementing its associated value in each row by 1, while querying it consists of taking the *minimum* of each associated value. The conservative update limits the increment to only those entries

that equal the minimum value. The fundamental idea is that if a collision of 2 strings under 1 hash function is rare, the simultaneous collision of the same 2 strings under 2 independent hash functions is extremely rare. Thus, with additional rows, it becomes increasingly unlikely that the minimum value assigned to an n-gram will over-estimate the true count of its occurrence.

The hash collision probability (or rather, the probability of colliding across all hash functions) is given in the original (Cormode & Muthukrishnan, 2011) paper, and is philosophically just a generalized birthday problem calculation. The more complicated calculation (and more pertinent) is, however, the probability that the count associated with a string is incorrect (too high). This could occur from a second string colliding across all the hash tables, but more likely from multiple distinct strings each individually colliding on only a few hash tables, but collectively resulting in a net increase over all the entries for the original string. That this probability is bounded and easily estimated is what makes (Cormode & Muthukrishnan, 2011) so notable here. Below we present their confidence interval bounds. Matching intuition, increasing the number of entries per hash table achieves 1st order reduction in the width of the confidence interval, while increasing the number of hash functions increases our confidence as $1-\mathrm{e}^{-d}$.

The memory utilization of this approach offers a significant advantage —the storage of a string requires 1 byte per character, thus $(n+1)$-grams will take more space than n-grams in memory. However, the Count-Min-Sketch offers a fixed memory data structure —assuming 4 bytes per entry (a 32-bit integer), it will consume $4 \times w \times d$ bytes in memory *independent* of the dataset being studied. When used to estimate probabilities (rather than integer counts), we have the confidence interval

$$Pr\left[p_i \leq query(\omega_i)/N < p_i + \varepsilon\right] > 1 - \mathrm{e}^{-d}, \tag{4.2}$$

where N is the number of items stored (including repeats), p_i is the true empirical frequency of ω_i and $\varepsilon \approx \mathrm{e}/w \propto 1/w$ (Cormode & Muthukrishnan, 2005).

We implemented this algorithm in `C++` using a reference MurmurHash3 hash implementation coupled with a pair-wise hashing optimization (Kirsch & Mitzenmacher, 2006), and then exported our class to `R` via `Rcpp` (Eddelbuettel et al., 2011). Writing the core implementation in `C++` allows for precise, efficient, and predictable memory utilization, while the `Rcpp` binding allows for its convenient use via the `R` scripting language.

4.2.1 Usage

The `cmscu` library has only a few methods that wrap the entire functionality. We describe the 3 primary methods below, and refer the reader to the GitHub page[1] for the full documentation and package.

Sample usage is given below:

```
dict <- new(FrequencyDictionary, 4, 10^6);
# 4 is the number of hash functions(d) and 10^6 is the width(w)
# Total size (in bytes) = 4 x w x d

bigrams <- c('this is', 'is sample', 'sample usage');
dict$store(bigrams);

test <- c('this is', 'not present', 'sample usage', 'this is');
counts <- dict$query(test);
# counts is c(1,0,1,1)
```

[1] `http://www.github.com/jasonkdavis/r-cmscu`

Table 4.1: Methods for `cmscu`

`dict <- new(FrequencyDictionary, d, w)}`	Initialize a dictionary with `d` rows ($d = 4$ gives $> 98\%$ confidence) and w bins
`dict$store('string')`	Update the frequency count associated with string `'string'`
`dict$store(c('s1', 's2', ...))`	Update the counts associated with all of the strings simultaneously
`dict$query('string')`	Query the (approximate) frequency count associated with `'string'`
`dict$query(c('s1', 's2', ...), n=1)`	Query the counts of all the strings simultaneously over `n` OpenMP threads (if available)

4.2.2 Comparison

We compare the application of our library to that of the `tm` package (Meyer, Hornik, & Feinerer, 2008), a common text-mining package in R. This package is frequently recommended for the analysis of n-grams in `R` specifically for its `DocumentTermMatrix` class, which counts the occurrences of each string-type per document and stores the integer values into a large matrix. It is this functionality that `cmscu` specifically offers an efficient alternative to —`tm` also provides high quality text-processing utilities that we do not attempt to replicate.

Perhaps unsurprisingly, by tailoring our data structure to the problem at hand, `cmscu` outperforms `tm`'s `DocumentTermMatrix` by orders of magnitude in the task of n-gram frequency analysis and information density calculation. The reason for this is that the `DocumentTermMatrix` solves a more general problem, which requires the storage and organization of data unnecessary for our calculation and prevents a linear scaling in the size of the corpus. While it is a powerful tool for document classification, and bundles useful functionality for cleaning and preparing raw strings, it is not optimal for the specific task of n-gram information density computation (despite its usefulness for such tasks with much smaller data sets).

We benchmark the creation, initialization, and evaluation of increasingly large datasets (the first k lines from the Yelp Inc. dataset) averaged over 10 runs in Figure 4.1. We run our CMS-CU implementation with 4 rows of 2^{24} entries, using a fixed 1GB of RAM for each run. For small data sets, the cost associated with this unnecessarily large memory allocation outweigh the `tm` calculation; however, as we approach 10^4 lines of the Yelp dataset, we are 18 times faster in total run-time. For larger sizes, we were unable to even run the reference `tm` code due to the non-linear scaling in its algorithmic complexity and memory use.

For a dataset of k lines, we compute k values with the `tm` result being exact and our CMS-CU result being approximate. We compute the root-mean-squared (RMS) error of our calculation in Table 4.2, where

$$RMS(x,y) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_i - y_i)^2}. \tag{4.3}$$

Table 4.2: Root-Mean-Squared Error of CMS-CU Output over Increasingly Large Datasets

k	10^1	10^2	10^3	10^4
RMS	2.22×10^{-16}	4.31×10^{-16}	7.30×10^{-3}	8.75×10^{-3}

When $k = 10^1$ and $k = 10^2$, there is no difference (machine precision) in the output

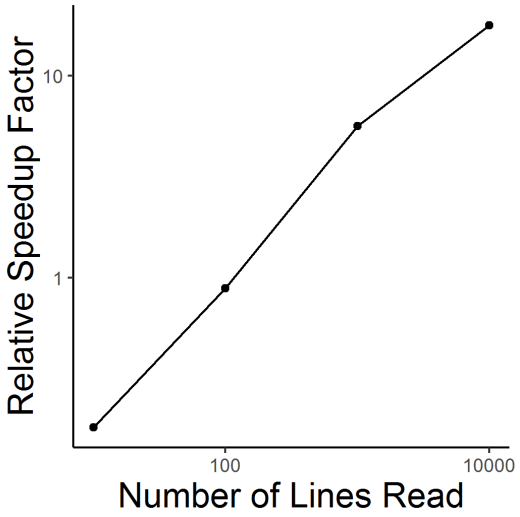


Figure 4.1: log-log plot of the calculation time of `tm` relative to the calculation time of our package, averaged over 10 runs

Due to the nonlinear scaling of `tm`, our implementation becomes increasingly faster relative to `tm` as the dataset increases in size. The memory requirements of `tm` prevented the comparison of larger data sets.

due to our large allocation which suggests the absence of any hashing collisions. As k increases, collisions expectedly occur but the average information density, $-\frac{1}{n}\sum_{i=1}^{n} log_2 p(w_i)$, of a review maintains a precision of 10^{-3}. This could be improved by using additional rows in the CMS-CU structure (going from 4 to 6, for example).

4.3 Information-Theoretic Structure of Yelp Reviews

The development of efficient n-gram storage and querying techniques affords analyzing text using more sophisticated algorithms. Standard practice involves training a model on held out data and using it to predict test data. However, it is almost always the case that unseen n-grams will occur in test data, resulting in poor model performance. Applying smoothing techniques adjusts n-gram probability estimates to help account for missing data and increase model performance (i.e, decrease model perplexity)[2]. The term 'smoothing' comes from the fact that these algorithms tend to make the distribution more uniform by adjusting low probabilities upward and high probabilities downward (Chen & Goodman, 1999).

To date, the most accurate models estimate the maximum likelihood of an n-gram using both higher —and lower-order n-grams. Two types of models exist: (1) *Back-off* models use lower order n-grams (such as bigrams and unigrams) to estimate the maximum-likelihood of higher order n-grams (such as trigrams), but only when data from the higher-order n-gram is missing. Thus it *backs off* to the lower-order $(n-1)$-gram until the value is defined. (2) *Interpolated* models use estimates from lower-order n-grams even when higher-order n-grams are defined. Interpolated models are defined recursively as a linear interpolation between the nth-order maximum likelihood model and the $(n-1)$-th order smoothed model (Chen & Good-

[2]The accuracy of specific Information-Theoretic models on estimating unseen data that vary in the length of n or the complexity of the algorithm can be determined by measuring its cross-entropy or more specifically model *perplexity* (Bahl et al., 1983; Jelinek, Mercer, Bahl, & Baker, 1977; Jurafsky & Martin, 2000).

man, 1999, p. 364). The most accurate interpolation models penalize higher —and lower-order n-grams using a *discount* parameter for higher n-grams and a *smoothing* parameter (often defined in part by the count of the higher-order n-gram) on lower order n-grams. It is reasonable to use both lower and higher-order n-grams together when estimating higher-order n-grams because the frequency of the $(n-1)$-gram will typically correlate with the n-gram and has the advantage of being estimated from more data. For this reason, such models tend to accurately estimate unseen data. Crucially, both types of models use lower-order n-grams to estimate *missing* higher-order n-grams (essentially interpolated models are back-off models when n-grams are missing). Similarly, in the occurrence of zero count data, where w_i is never seen (at any n- or $(n-1)$-gram) the standard procedure across both models, and one we adopt here, is to estimate its value via a uniform distribution, $p_0(w_i) = 1/|v|$, where $|v|$ is the model's vocabulary (the number of unique 1-grams in the training data).

At present, the most accurate smoothing technique is a variation on what is known as Kneser-Ney smoothing (Kneser & Ney, 1995). What makes Kneser-Ney smoothing more accurate than other smoothing techniques is how it estimates unigrams (1-grams). While other models simply take the frequency of the unigram, Kneser-Ney smoothing estimates this value based off how likely w_i is to occur in an unfamiliar context:

$$p_{KN}(w_i) = \frac{|\{w_i : 0 < c(w_{i-1}, w_i)\}|}{|\{(w_{i-1}, w_i) : 0 < c(w_{i-1}, w_i)\}|}, \tag{4.4}$$

where $|\{w_i : 0 < c(w_{i-1}, w_i)\}|$ is the total number of bigrams w_i completes divided by the total number of unique bigrams, $|\{(w_{i-1}, w_i) : 0 < c(w_{i-1}, w_i)\}|$. A common example used to explain why this is a better estimate of higher-order n-grams is to consider the unigram 'Francisco'. Suppose 'Francisco' occurs frequently throughout our dataset. However, it only ever occurs after the word 'San'. Because it is unlikely to occur in any other bigram context, its unigram probability should not be high as this would result in an inflated probability estimate for 'Francisco'.

We use a version of Kneser-Ney smoothing, *interpolated Fixed Modified Kneser-Ney* (iFix-MKN), to estimate conditional trigram and bigram probabilities (Maximum Likelihood Estimation). Our model is a direct implementation of that found in (Chen & Goodman, 1999, section 3). Kneser-Ney smoothing uses both a discounting parameter which subtracts some value from non-zero count n-grams, and a smoothing parameter that maximizes the probability of obtaining an accurate estimate from lower-order n-grams. Crucially, our implementation is *fixed* in that both the discounting parameter and smoothing parameters are defined prior to applying the model to test data. Both parameters can be more accurately defined when estimated using held out data prior to analysis. While models that estimate their parameters outperform their fixed counterparts, iFix-MKN outperforms all other smoothing techniques (including those that use held out data to estimate parameters), with the exception of the version of MKN that uses estimated parameters. As a result, iFix-MKN has the benefit of being both simpler and more domain-general. In what follows we leverage an implementation of iFix-MKN using our `cmscu` package to explore and quantify language use in a large natural dataset.

4.3.1 Current Study

Our analyses are motivated by recent studies that show a message's Information-Theoretic structure is influenced at a variety of linguistic levels including syntactic variation and phonetic reduction (Aylett, 1999; Genzel & Charniak, 2002; Aylett & Turk, 2006, 2004; Levy & Jaeger, 2006; Jaeger, 2010; Mahowald et al., 2013). Specifically, the amount of information present across a message is shown to increase over time, abiding by the *entropy rate constancy* principle —a message's information density increases at a stable rate (Genzel & Charniak, 2002) —perhaps in an effort to provide relevant content against channel noise. That is, the

cognitive agent works to structure their utterance in a way that maintains the highest rate of information without breaking channel capacity (determined by the amount of noise in the system). When at risk of breaking channel capacity, language users might add optional low-information words to high information-dense messages (Jaeger, 2010), or simply slow down their utterance (Aylett & Turk, 2006, 2004) effectively spreading an otherwise information-dense message over a longer utterance. These findings suggest that language users are sensitive to channel noise and adjust their utterances to match the channel's capacity in an effort to balance redundancy with confidence in signal transmission. To do this the theory of Uniform Information Density (UID), building off previous work, e.g., the smooth signal redundancy hypothesis: (Aylett & Turk, 2004), suggests language users try to avoid "peaks and troughs" in information density. As a result, language users exhibit an inverse relationship between language redundancy and predictability in an effort to communicate efficiently.

Theories such as UID posit that the information density of one's message may be attuned to the expectations the producer holds about their intended audience. One possibility is that a more established common ground due to a larger number of shared experiences may result in a decrease in channel noise, due to a shared common ground, that afford more complex information-dense language use (H. H. Clark & Brennan, 1991). Indeed, many studies have shown that language users make assumptions about their audience and structure their own utterances with these assumptions in mind (Brennan & Williams, 1995; Jaeger, 2013; Krauss & Fussell, 1990; Pate & Goldwater, 2015). For example, the information density of a language user's Yelp reviews is higher when their network of friends is more densely interconnected (Vinson & Dale, 2016). Similarly, microblog posts on Twitter about specific events, such as a baseball game series, are more information-dense toward the end of a sporting event than during (Doyle & Frank, 2015). Such findings add to growing evidence that language users are sensitive to the knowledge they share with their audience. This sensitivity is reflected in how they structure their utterances.

Method Few studies show directly how one's audience perceives the helpfulness of a producer's language use. In this example application of `cmscu`, we explore in what ways the information density of a Yelp user's review, estimated via iFix-MKN, can predict how useful, funny and/or cool (U/F/C) it is to its reader. The `cmscu` package greatly facilitates these exploratory analyses of natural data sets. With it we are able to efficiently process massive amounts of text from the Yelp, Inc. Dataset Challenge corpus using one of the most sophisticated algorithms to date.

We estimated a total of five measures from Yelp review text: two information density measures and two uniformity measures as well as review length. We even estimate information measures over trigrams for this analysis, which `cmscu` permits with great flexibility. As this section serves only as an example analysis, we detail these measures in Appendix I4.6 below. The measures are based on analysis of bigrams, and trigrams using the iFix-MKN algorithm requiring estimation of conditional probabilities in Yelp text to the second order (trigrams). These measures are summarized briefly in Table 4.3.

Table 4.3: Summary of Information-Theoretic measures

n	Density Measure	Uniformity Measure
2	Average (iFix-MKN) Bigram Information: ABI,	Variance of Bigram Information: $\sigma(BI)$
3	Average (iFix-MKN) Trigram Information: ATI,	Variance of Trigram Information: $\sigma(TI)$

The standard deviation of each information density measure was taken as the measure

of variance (inverse uniformity). In addition to the four measures in Table 4.3, we included review length, giving us five variables used to predict U/F/C. Increasing n quantifies a successively more multiword estimation of information density and uniformity in language use. Given the nature of iFix-MKN, such that in the event of missing n-grams the model backs off to a lower-order n-gram, in our analysis of the Yelp dataset the correlation between trigram information and bigram information is high, r=.93, t(1.03 mil) = 2527, $p <$ 2.2e-16. For this reason, we predict U/F/C using two separate models, one including bigram information/variance and length and the other including trigram information/variance and length. Information measures become more and more sparse as we increase model complexity and so there is a trade-off in using different models. Estimating higher-order n-grams gives a better measure of the actual structure of the language, while estimating lower n-grams provides a better model fit. Moreover, when modeling real language use, such as that from the Yelp community, it can vary widely from one review to the next leaving even very well trained models weak predictors. In such cases, estimating lower-order n-grams may prove to be a stronger predictor of behavioral phenomena. Finally, our variance measures will correlate with information density measures due to the presence of a true zero in information density. This will inflate the variance accounted for within our regression models. We adjust for both issues by first predicting each uniformity measure by its information density measure using linear regression (`lm`) in `R`, and then taking the residual of that model as our true estimate of information uniformity; $r\sigma(BI)$, $r\sigma(TI)$

Higher-order n-gram models are more computationally expensive to estimate using iFix-MKN. However with `cmscu` it is relatively straightforward and easy to deploy in `R`. Though the precise n-gram model details are outside the scope of this example application of `cmscu`, we offer the reader a breakdown of our variables and modeling approach in Appendix I. This will also serve as an example strategy for deploying `cmscu` in more detail.

Though simply an example application, prior research motivates some predictions: (1) *Information density will be positively related to U/F/C ratings.* The language use within reviews is most likely already abiding by the constraints of its community. For this reason, information-dense messages should hover closer to the channel's capacity, thus providing more helpful content without too much risk of being misunderstood. (2) *Information variance will be positively related to U/F/C.* Increased variance may be associated with higher reader ratings. Specifically, more variance in information across a message may be related to higher reader ratings as it may suggest reviewer's are inserting low information-dense words to lower the risk of presenting information that may be misunderstood.

Results In all cases, a negative binomial model predicted, perhaps surprisingly, a large portion of variance. Table 4.4 and Table 4.5 report the 95% Confidence Intervals (CI_β) and associated Z-scores for each predictor variable as well as the overall R^2_{adj} for each model containing bigram and trigram measures respectively. There were no differences between the trends in either bigram or trigram model predictors. That is, all measures were highly significant positive predictors of Useful (Figure 4.2) Funny (Figure 4.3) and Coolness (Figure 4.4) ratings, such that an increase in review length, information density and variance increased the probability of a review receiving more U/F/C ratings. In order, review length (*log-Length*) was the strongest predictor followed by information density measures (*ABI* and *ATI*) and last the variance of information density, *rσ(BI)* and *rσ(TI)*. Because there was little difference between trigram and bigram models and because trigrams are considered a stronger estimate of the structure of the language itself, Figures 2-4 present only the trigram model's predicted U/F/C ratings by (A) *log-Length*, (B) *Average Trigram Information* and (C) *Variance of Trigram Information* respectively.

Overall this exploratory analysis suggests that the composition of language use, in terms of Information-Theoretic structure, significantly predicts the impression of a review from its readers. These exploratory analyses open up interesting avenues for future research on language use. For example, showing that the variance of information is related to reader ratings

Table 4.4: Bigram by Reader Rating Negative Binomial Model

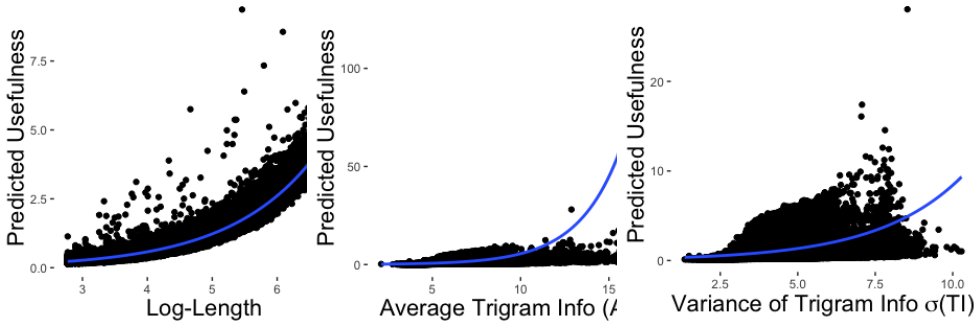
Rating type	Predictor	95%(CI_β)	Z-value	Effect Size
Useful	*Intercept*	(-.169, -.163)	-104.4*	R^2_{adj} = .15
	log-Length	(.585, .591)	367.0*	
	ABI	(.175, .182)	113.3*	
	$r\sigma(BI)$	(.040, .043)	24.5*	
Funny	*Intercept*	(-1.121, -1.111)	-425.6*	R^2_{adj} = .09
	log-Length	(.647, .657)	254.1*	
	ABI	(.355, .365)	143.6*	
	$r\sigma(BI)$	(.075, .086)	30.8*	
Cool	*Intercept*	(-.828, -.819)	-365.8*	R^2_{adj} = .09
	log-Length	(.616, .625)	275.5*	
	ABI	(.197, .206)	90.3*	
	$r\sigma(BI)$	(.102, .113)	46.7*	

*`p < 2e-16`, *r is the residual from* `lm` *models described in text*

Table 4.5: Trigram by Reader Rating Negative Binomial Model

Rating type	Predictor	95%(CI_β)	Z-value	Effect Size
Useful	*Intercept*	(-.167, -.161)	-102.9*	R^2_{adj} = .15
	log-Length	(.590, .597)	371.7*	
	ATI	(.161, .167)	102.4*	
	$r\sigma(TI)$	(.043, .050)	28.5*	
Funny	*Intercept*	(-1.118, -1.108)	-424.4*	R^2_{adj} = .09
	log-Length	(.653, .663)	256.2*	
	ATI	(.352, .362)	140.8*	
	$r\sigma(TI)$	(.062, .073)	25.8*	
Cool	*Intercept*	(-.825, -.816)	-364.4*	R^2_{adj} = .09
	log-Length	(.625, .634)	278.8*	
	ATI	(.175, .183)	79.3*	
	$r\sigma(TI)$	(.114, .123)	51.4*	

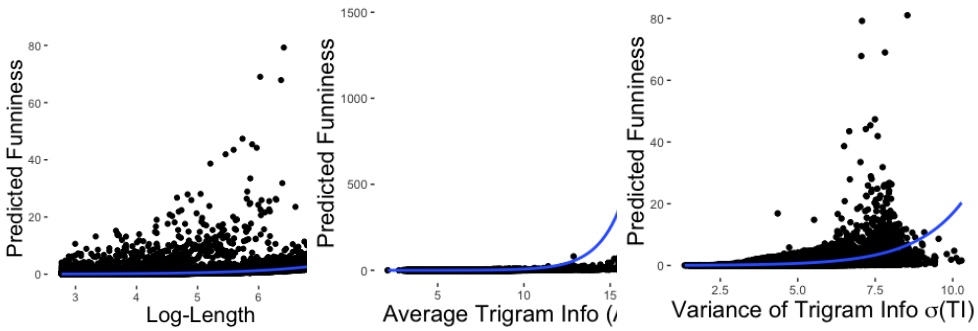
*`p < 2e-16`, *r is the residual from* `lm` *models described in text*



(a) (b) (c)

Figure 4.2: Predicted Usefulness ratings by (a) log-Length, (b) Average Trigram Information (ATI) and (c) Variance of Trigram Information $\sigma(TI)$.

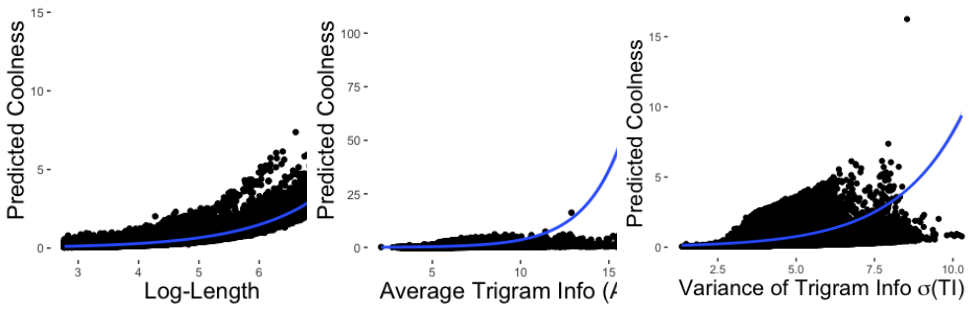
Predictions are provided by the full negative binomial model controlling for other variables.



(a) (b) (c)

Figure 4.3: Predicted Funniness ratings by (a) log-Length, (b) Average Trigram Information (ATI) and (c) Variance of Trigram Information $\sigma(TI)$.

Predictions are provided by the full negative binomial model controlling for other variables.



(a) (b) (c)

Figure 4.4: Predicted Coolness ratings by (a) log-Length, (b) Average Trigram Information (ATI) and (c) Variance of Trigram Information $\sigma(TI)$.

Predictions are provided by the full negative binomial model controlling for other variables.

as well as information density measures may suggest readers are more likely to find both high information and decreased channel noise useful for comprehension. Future research interested in understanding what aspects of online reviews readers find useful, funny or cool should be sure to investigate various other linguistic features, such as simpler lexical-level variables (e.g., curse words), that might be highly correlated with more sophisticated information measures presented here. This is outside of the scope of the current demonstration, but the tool we introduce here may permit such analyses efficiently.

4.4 General Discussion

Efficiently processing large amounts of textual data is a problem at the forefront of behavioral science. One way to advance this process, as we demonstrate here, is by adapting well-known tools from computer science by developing statistical packages in programs used by behavioral scientists. This effectively broadens the number of tools that can be used by behavioral scientists while redefining the problem space wherein that tool is applicable. Here we use a sketch algorithm known for its efficiency in processing massive real-time data (Cormode & Muthukrishnan, 2011) to approximate the information density of words, using a sophisticated smoothing algorithm (iFix-MKN), across millions of online reviews. We show it to be a successful tool toward discovering interesting behavioral phenomena.

Our analysis shows the `cmscu` package can be used to process *n*-gram data at speeds and scales beyond the reach of commonly available `R` packages. In real-world applications, such as our Yelp data analysis, which used a modified Kneser-Ney implementation powered by the `cmscu` library, was able to process 5% of the 2.2 million reviews we evaluate in under a hour, on a single core, on commodity hardware. Our analysis was configured to evaluate up to quadgrams, requiring 8 separate `cmscu` instances each configured to occupy 1gb of RAM. A much earlier (attempted) analysis on the dataset, built upon `tm` using its `DocumentTermMatrix` object, had run for over 2 months without finishing —the algorithmic and memory scaling requirements resulted in constant swapping of hard drive space, which effectively brought the computations to a standstill. Though limited in features compared to `tm`, `cmscu` is simple to use, requires few lines of code, and scales with the processing power of one's computer and size of one's dataset. That is, users may specify memory usage *a priori*, in line with their hardware's capabilities, and nonetheless obtain a useful analyses, independent of the size of the dataset under study.

Using `cmscu` we explore possible relationships among reader ratings and sophisticated estimations of Information-Theoretic structures of review text in a large Yelp, Inc. dataset, a process difficult for common scripting packages. Indeed, the sheer size of our dataset affords the possible discovery of subtle, but interesting relationships such as those between the linguistic choices of language users and their audience. For this reason our predictions are inherently exploratory. Though our predictions are broad, and somewhat intuitive perhaps, our findings build on previous work that shows language users structure their utterances with their intended audience in mind (Jaeger, 2013; H. H. Clark & Brennan, 1991; Brennan & Williams, 1995).

We speculate that reader ratings mark the successful transmission of some useful information by its producer. If so, longer reviews, those with more information density and those with greater variance in information are more likely to be successfully transmitted. Interestingly, one possibility is that higher reader ratings are more probable for reviews that are more information-dense and more variable because reviewers insert low information-dense words in higher information-dense messages which may help to facilitate comprehension (Jaeger, 2010).

Another possibility worth further exploration is whether simpler linguistic factors might account for reviewer ratings more so than relatively sophisticated information measures. Specific lexical words provide a window into the semantic content that might be influencing reader ratings. This is inherently different than Information-Theoretic measures which target how likely those same words are to occur given the frequency of its context. Future studies might include both semantic and Information-Theoretic factors to determine what aspects of one's language use are considered more helpful to its audience.

These exploratory findings add to ongoing research throughout the behavioral sciences. However, further investigation is needed in order to assess the generality and accuracy of our findings. Our analysis explores new avenues for fruitful scientific data exploration. Such is the case with many exploratory studies. Yet, the exploration of interesting behavioral relationships is often inaccessible to classically trained behavioral scientists whose collection methods are often guided, justifiably, by detailed specific theoretical concerns.

4.5 Conclusion

Current problems in science and industry involve processing increasingly large and complex data sets which necessitates the development of novel scalable computational tools. Mathematical and computational scientists are naturally a good fit to discover such solutions as their training is often geared toward finding solutions to engineering problems, paying little mind to interesting behavioral phenomena discoverable at their fingertips. Yet, more and more freely available data sets such as those from the Netflix challenge, Yelp Dataset Challenge as well as Twitter's API and Yahoo!'s release of 100 million images, are overwhelmingly loaded with interesting behavioral nuances that can be harnessed to answer longstanding questions in the behavioral sciences and also increase the success of newly developing machine learning algorithms that aim to predict future behavior. Indeed, some cognitive scientists argue we are currently in the midst of a revolution: "to take back behavioral data, and —just as in the last cognitive revolution —to demonstrate the value of postulating a mind between browsing history and mouse movements" (Griffiths, 2015).

Unfortunately, taking back behaviorally relevant data is not always so straightforward. Even after obtaining the data many behavioral scientists do not know what tools are necessary to address their questions. Even when the most cutting-edge tools are freely available, few behavioral scientists are trained in methods that could enable them to harness these powerful tools. To conclude, we return to our three key methodological observations summarizing how they were successfully used to facilitate this process.

First, We described how "sketch" techniques help process large amounts of data efficiently and argue these are critical for the coming 'big data' age in cognitive science (Griffiths,

2015). This provides one instance that shows it is possible that behavioral scientists can often find more efficient data techniques for their problems (Section 2, `R`-package `cmscu`). Second, corpus or other data analysis can make ready use of these specialized solutions (Section 3, Information-Theoretic Structure of Yelp Reviews). We demonstrate a fruitful domain in which such a strategy can apply —the Information-Theoretic analysis of language structure in corpora. We then used our library to implement a sophisticated *n*-gram analysis (iFix-MKN) to explore the statistical properties of language that predict successful communication. Finally, multidisciplinary cross-fertilization is crucial (General Discussion). We revisited important messages about cross-disciplinary interactions and suggest that adopting a wide range of tools for various disciplines can help to ensure that behavioral scientists find efficient solutions for their problems, while giving computational scientists and engineers exciting behavioral problems to apply their research to.

Now, more than ever, interdisciplinary collaborations among behavioral and computational scientists are required in order to successfully accomplish our goals. To this end, some argue we are in the midst of a new era of scientist —the computational social scientist —who's focus lies at the intersection of cognitive and computer science (Lazer et al., 2009). This manuscript belays the success of one such method, interdisciplinary collaboration, that can help maintain the pace of scientific discovery within the behavioral sciences. Our work fits within the broader theme of advancing discovery across the sciences illustrating that interdisciplinary collaboration —which can connect previously intractable problems with new tools and methods —is a successful approach toward accomplishing this goal.

4.6 Appendix I

The two information density measures are the Average Bigram Information (*ABI*) and the Average Trigram Information (*ATI*):

$$ABI_j = \frac{1}{N-1}\sum_{i=2}^{N} -\log_2 P_{iFix-MKN}(w_i|w_{i-1}), \tag{4.5}$$

$$ATI_j = \frac{1}{N-2}\sum_{i=3}^{N} -\log_2 P_{iFix-MKN}(w_i|w_{i-1},w_{i-2}) \tag{4.6}$$

where N is the number of words in the j_{th} review, $P_{iFix-MKN}(w)$ is the iFix-MKN probability of word w, and w_i is the ith word. Bigram Information is the probability of w_i given w_{i-1} and Trigram Information is the probability of w_i given the joint probability of w_{i-1} and w_{i-2}. We then obtain the information density of a review by taking the average amount of information across the review. We calculate the standard deviation, σ, of each review's average information by each *n*-gram and take this as an inverse measure of a message's uniformity: $\sigma(BI)$, and $\sigma(TI)$. An example implementation of iFix-MKN can be found on the websitefootnote 1 and specific details about the algorithm can be found in (Chen & Goodman, 1999, section 3).

We estimate both bigram and trigram information, even though estimates within the trigram model rely, in part, on bigram model estimates. We do this because Maximum Likelihood bigram models will typically account for more unseen data, while trigram models are a better measure of the language's actual structure (e.g., afford better word predictability when the data exists). We avoid making the decision to balance this trade off by modeling both explicitly in two distinct models and comparing those models to one another.

Independent Variables We trained the information models on Yelp reviews from the United States only to avoid training on non-English reviews (though future research might look to see if different languages show similar results to those found here). Half of the Yelp, Inc. dataset, 1.1 million reviews, was used to train each model which was tested on the remaining half.

After obtaining information density and uniformity measures for each review within the test set we removed reviews shorter than 15 words in length (though various techniques to control for *n*-gram reliablity do exist (Martin & Jurafsky, 2000; Frank & Jaeger, 2008; Vinson & Dale, 2014b)). This reduced the total number of reviews from 1.1 million to 1.03 million ($< 5\%$ reduction).

We anticipated the possibly of an inflated variance due to multicollinearity when including both information density and uniformity measures within the same model. The variance (inverse uniformity) naturally increases as information density increases, due to the presence of a true zero. For this reason, we took the residual of each uniformity measure, first predicted by its respective information density measure as the true measure of uniformity ($r\sigma(BI)$ and $r\sigma(TI)$). After, a variance inflation factor (`VIF`) analysis (Craney & Surles, 2002; Stine, 1995) in `R` (Library `CAR`) was used to determine whether the new predictor variables exhibit collinearity with any other predictor variable. None of our predictor variables showed signs of strong collinearity ($\texttt{VIF} < 2$). All variables were then centered and standardized for the purpose of interpretation.

Assessing Model Fit U/F/C ratings are count variables associated with each distinct review. Because of this we initially use a Poisson regression model in `R` to predict each rating. We compare this model against a null intercept model using a chi-squared test of difference of log-likelihoods. We found the full Poisson model was an improvement over the null model; however, the variance of each reader rating was greater than the mean (*Useful*: $M = 1.05$, $SD = 2.17$, *Funny*: $M = .46$, $SD = 1.60$, *Cool*: $M = .56$, $SD = 1.72$) indicating possible over-dispersion (larger number of zeros). One assumption within Poisson models is that the variance equals the mean. Violating this assumption may result in underestimated standard errors or inflation in model significance. Because of this, the distribution might be better fit by another model that does not make this assumption (Long, 1997). Specifically, we compare the Poisson regression against a negative binomial regression that does not require the variance equal the mean (thus adjusting for over-dispersion)[3]. Again using a chi-squared test on the difference between log-likelihoods, we found the negative binomial model was an improvement over the Poisson model for all (U/F/C) ratings. Crucially, the Poisson model results revealed the same exact trend (similar significant predictors) as the results from the negative binomial model, but with higher coefficient estimates. This suggests the negative binomial model is in fact adjusting for inflation of variance present in the Poisson models. Thus, we report the results from the negative binomial models.

To assess the significance of the negative binomial model, we use a likelihood ratio test comparing the deviance of a null model —predicting U/F/C using the intercept only —and the full model. We found the negative binomial model is significantly better at predicting U/F/C than the null model. Importantly, the negative binomial regression model does not allow for a straight forward interpretation of effect size. Taking after previous research, we use an adjusted likelihood-ratio-index (Abney, Gann, Heutte, & Matlock, Submitted; Long & Freese, 2006) a type of pseudo R^2_{adj} (Mittlböck, Schemper, et al., 1996) as our measure of effect size:

$$R^2 = 1 - (L_{\text{fitted}}/L_{\text{intercept}})^{2/n} \tag{4.7}$$

Where L is the data likelihood.

[3]Other possible models include the zero inflated Poisson model and zero inflated negative binomial model. Both models assume that over-dispersion is, in part, due to another process that can be modeled independently. One possibility is that certain reviews were simply never read. Though a clearly discernible independent variable, there is no such variable within the dataset that could be used to model this process using a zero inflated-Poisson or negative binomial.

Chapter 5

Statistics in the Fingertips: Typing Reveals Word Predictability

5.1 Introduction

Typing can serve as a window into cognitive processes underlying language planning and its connection to motor action (Rumelhart & Norman, 1982). Typing provides an array of rich reaction-time data during planning and execution, and thus provides a kind of "microstructure" of language performance. In the past three decades, much has been learned about expert typing and how it relates to cognition. Usually these effects are demonstrated in "copyist" tasks, rather than spontaneous composition (for early reviews see Cooper, 1982; Gentner, Larochelle, & Grudin, 1988) thus researchers must infer that their findings translate to more naturalistic behavior. Despite this progress, *spontaneous* and *natural* language production has not been extensively studied in typing. Even now when new tools provide a means of collecting spontaneous typing from the comfort of one's own computer, copyist tasks remain the primary method for determining whether language planning influences production at the fingertips (Cerni, Velay, Alario, Vaugoyeau, & Longcamp, 2016; Pinet et al., 2015, 2016; Scaltritti, Arfé, Torrance, & Peressotti, 2016). In this study, we show that online tools for data collection permit rapid generation of large and natural typing corpora. We combine these millisecond-level data with current methods for measuring word predictability from "big data" (Behmer & Crump, 2015). Our results show that the dynamics of natural everyday action —at the fingertips —echo single- and multi-word statistics.

Our work is inspired by other work that has begun to integrate dynamics of action with cognitive tasks to explore how a cognitive process unfolds in time, and how action and cognition relate during mental processing (Spivey & Dale, 2006; Spivey, Grosjean, & Knoblich, 2005). Such measures have included computer-mouse tracking (J. Freeman, Dale, & Farmer, 2011), eye tracking (Tanenhaus et al., 1995), choice dynamics (Ross, Wang, Kramer, Simons, & Crowell, 2007), speech (Kawamoto, Kello, Jones, & Bame, 1998; Kawamoto, Kello, Higareda, & Vu, 1999), muscle activations (Moreno, Stepp, & Turvey, 2011), and more. Keystroke dynamics provide an ecologically valid sequence of reaction times. These reaction times occur at different levels, such as the onset of a word and the patterns of keystrokes within that word. As we review below, much recent work has already shown the value of typing as a source of latency data.

Language may rely on prediction as a core cognitive process (for recent review see Huettig, 2015; Van Petten & Luka, 2012; Shaffer, 1973), as may all of the human brain itself (Friston, 2010). Probabilistic models of cognitive prediction have been powerful in a variety of domains (see A. Clark, 2013, for review), and they find fruitful application in language, too. Over the past decade researchers have shown the predictability of one's message in con-

text influences language production in what appears to be an effort to maintain a stable rate of statistical predictability (Aylett & Turk, 2006, 2004; Tily et al., 2009; Mahowald et al., 2013; Levy & Jaeger, 2006; Frank & Jaeger, 2008; Genzel & Charniak, 2002; Jaeger, 2013, 2010). Specifically, some interest surrounds the use of function words, showing that when a message is higher in information density (e.g., lower in predictability) one's speech rate may be slower (Aylett & Turk, 2004) and may include optional relativizers (e.g., the use of the word "that" when it's optional) (Jaeger, 2010). Indeed function words may even serve as predictive frames for content words. Because they participate in a message's overall statistical structure, they help categorize content words by part of speech (Monaghan, Chater, & Christiansen, 2005; Mintz, Newport, & Bever, 2002; Valian & Coulson, 1988; Gibson et al., 2013; Piantadosi et al., 2011). These findings suggest the language production system is organized to achieve efficient communication (Jaeger, 2013). It does so by providing a *balanced* message —robust (i.e., predictable) but information-dense. To do this it must unfold adaptively over time, responding to its predictability in context.

Recently, Smith and Levy (2013) developed a probabilistic model of reading time that is based on a logarithmic relationship between word predictability and processing speed. Specifically, they show that the probability of a given word, p_{word}, is related to processing time in the following way:

$$\text{total processing time} \approx -\theta log p_{word} \tag{5.1}$$

In practice, p_{word} can be estimated by a variety of measures, such as conditioning word predictability by prior word (Saffran, Newport, & Aslin, 1996) or calculating its contextual diversity (Plummer, Perea, & Rayner, 2014) . However, this simple formula expresses a fundamental relationship between the brain's latency in processing a word and general predictability of that word in context. The basic insight, in sum, is that the mind uses the probabilistic sequential structure of language to guide processing.

This core idea makes important contributions to theories and models of the brain processes language. Probabilistic sequential structure has had great influence in well-known connectionist models of sentence processing (Elman, 1990), and more recently in theories of the cognitive architecture that supports linguistic interaction (Pickering & Garrod, 2013). The manner in which predictive processes work during language helps us to understand this complex human skill; these models may also inform a wide array of other tasks and theories in which prediction is featured prominently (e.g., see A. Clark, 2013).

Such findings bridge what we know about language to other cognitive processes, too. It is well known that this kind of logarithmic relationship holds in other cognitive tasks, such as Fitts' movement task (Fitts, 1954), and in decision making, such as in Hick's Law (Hick, 1952). Similar principles, linking information and performance in classic psychophysics, can be developed too for word-level language processing and production (for evidence that the units of typing exist at the word level, see Shaffer, 1973). Observations of such common principles have long been on offer in psychophysics (Stevens, 1957), and so it should perhaps not surprise us that such relations can be found readily in language processes. Linking language to these principles may help us to understand how language processing and production are founded upon the same core principles that govern other aspects of behavior.[1]

In this paper, we look to typing and show the same relationship Smith and Levy (2013) reveal: Typing time is related logarithmically to word predictability. We find that keystroke times have a relationship with word predictability, a level of analysis thought to reflect the units of skilled typing (Shaffer, 1973). Frequent content words are typed more quickly, even keystrokes internal to that word. In addition, when a function word is predicted by a prior typed word, its keystrokes are facilitated. This lends support to Smith and Levy's general observation

[1]Consider, for example, in vision: statistical variability in image presentation influenced picture naming latencies providing evidence that cognitive system is sensitive to probabilistic structures within one's environment (Bonin, Chalard, Méot, & Fayol, 2002).

about language processing, but we show that it holds in a form of large naturalistic language production across tens of thousands of keystrokes.

5.1.1 Lexical Effects in Typing

Logan and Crump (2011) argue that typing is underlain by two "loops", an outer loop that controls higher-level planning and word access, and an inner loop that executes sequences of keystrokes after word access. They argue that these processes are fundamentally encapsulated. If this is true, it also has implications for the overall language system. Though typing is possibly our most recently emerged skill related to language, the outer/inner loop dichotomy means that language planning and production interact only faintly, especially in a highly practiced task. The reason is that the interkey interval during typing is accounted for most prominently by the "digraph", the identity of the prior and current keys (Kandel, Peereman, Grosjacques, & Fayol, 2011). The relationship between interkey timing and word-level factors, such as lexical predictability, has long been regarded as a weak one. Research on typing expertise has found only weak effects, if any, of word-level variables over and above a digraph model (Pinet et al., 2015). Others have observed that word-level effects are likely present, but they probably contribute only a vanishingly small variance in explaining typing speed (Logan & Crump, 2011). Many previous studies haven't found effects beyond the word level on typing. This may have to do with the emphasis on copying. Word access effects may be more detectable at the keystroke level when keystrokes are generated voluntarily as part of a natural language production task. Here we quantify word predictability, like Smith and Levy (2013), by utilizing a large-scale corpora of balanced English text: The Corpus of Contemporary American English (Davies, 2009). We then collected a large-scale corpus of typing (cf. Cohen Priva, 2010), with the rare feature that participants do not simply copy words (still the most common method in typing research; Chukharev-Hudilainen, 2014) but rather contribute spontaneous production to a simple prompt. Under such conditions, despite the dominance of the digraph model (keystrokes are best predicted by the key-to-key biomechanics), we are able to detect a logarithmic relationship to word predictability at the *keystroke level*.

We show that information-theoretic measures based on word-to-word composition relate not only to overall word typing speed and first keystroke onset, but also to the typing speed *within words themselves*. The implication from spontaneous composition is that the dynamics of language production are directly impacted by the informational structure of the flowing text that is being composed. This recommends that cascade-based accounts of language production should participate in theories of linguistic performance (Kandel et al., 2011; Van Galen, 1991; Will, Nottbusch, & Weingarten, 2006; Roux, McKeeff, Grosjacques, Afonso, & Kandel, 2013; Scaltritti et al., 2016; Olive, 2014). In broader terms, we end with a discussion of how tasks like this one offer new ways to integrate information processing and dynamical systems accounts of language, typically seen as distinct traditions in the language sciences.

5.2 Experiment

5.2.1 Participants

A total of 317 participants were recruited through the online participant recruitment system at the University of California, Merced. Participants received course credit for their involvement. Participants were directed to a simple web page that included brief instructions and a textbox. They were asked to provide a detailed plot summary of their favorite movie. The textbox was equipped with JavaScript code to record each keystroke. After participants saved their text by pressing a button, they were debriefed on the same screen about the purpose of the study and provided researcher contact information should they have further questions.

Table 5.1: An illustration of integrating backspaces

key	raw IKI	integrated IKI
t	200 ms	<>
h	100	<>
e	200	<>
[backspace]	400	<>
[backspace]	100	<>
[backspace]	100	<>
a	100	1,200 ms

5.2.2 Procedures

Participants clicked on a URL which took them to a single page that provided a text box to enter their participant ID (for extra credit confirmation), the title of their favorite film, and a text box to provide a plot summary.[2] The text box was sized to accommodate approximately 100 words, and it did not permit use of the mouse. This allowed us to focus only on the stream of text. The text box was equipped with a keystroke logger using JavaScript, and an EventListener that ignored any mouse events inside the text box. This was to avoid having participants move their text cursor with the mouse, and thus disrupt the flow of text.[3] After piloting a version of our interface we found that it was best to minimize the impact on their flow of text, and simply ignore their attempt to use the mouse. Rather than let the participants know they should not use the mouse, such as an alert box or reminder, we simply forced the text cursor back to its most recent position. Once participants completed their plot summary, they clicked "Save" and returned to the online system to receive credit.

As participants typed, we tracked each keystroke by storing the ASCII character code. Using JavaScript's performance.now() function, we obtained a precise millisecond measure of each keydown event (Barnhoorn, Haasnoot, Bocanegra, & van Steenbergen, 2015). Participants were permitted to use the backspace to correct their text. We used a technique of "integrating" backspaces into the next keystroke's latency. For example, if a participant carried out the following sequence of keystrokes:

the[backspace][backspace][backspace]a dog

we would count the interkey intervals, *IKIs*, for each backspace (and each character deleted) as part of the keystroke timing for the determiner 'a.' Such backspace events would rarely be part of our data, because, as we describe further below, we used the recommended cutoff of 500ms as a marker of hesitation in typing (Chukharev-Hudilainen, 2014; Cohen Priva, 2010). For example, the latency on the keystroke for 'a' would not be 100ms, relative to the immediately prior backspace, but rather an integration of all events associated with prior backspaces, leading to a latency of 1,200ms. In this way we only focused on keystrokes that were not deleted and remained part of the final text, representing a participant's natural speed during production. The process can be further illustrated by the hypothetical data shown in 5.1.

5.2.3 Regression Models

As is common in such natural and spontaneous recording of behavior, we use a multiple, mixed effects regression using the statistical model *lmer* in *R*'s *lme4* package that is akin

[2]Readers may consult the interface at URL http://davevinson.com/exp/plot-summary

[3]If participants attempt to use their mouse and it disrupted their typing, it might have seemed unnatural to them. However it is likely that our data filtering procedures (see analysis, below) would detect such events as extreme latencies in interkey times, and would thus be omitted from analysis.

to an unbalanced repeated-measures setup —many keystrokes per participant, but each with a different cluster of words and character typing rate. We specified these two random effects, *words* and *subject*, in an effort to control for differences in words typed across participants and differences in variability across participants.[4]

As a conservative measure for making inferences, we chose a critical-p cut off of .001 to talk about a result as "significant," for all regression models outlined below. However, we take the liberty of proffering all regression model parameters in our results, including precise p values. As we specify below for each result, we used different regression models with transformed interkey time as our outcome variable, and combinations of the independent variables as predictors.

5.2.4 Dependent variable

A transformed latency of the interkey interval, z_{IKI} (i.e., within-subject z-score of each interkey interval) was used as the dependent variable for each keystroke. Each keystroke was assigned a reaction time by tracking the time, in milliseconds, after the last keystroke event. This was done for each and every keystroke, and backspaces were "integrated" into the next valid keystroke in the manner described above. We transformed interkey typing times within each participant to control for variability in typing speed, using a within-subject z-score over each participant's interkey times we standardize a given participant's keystrokes. In this way, each participant's latencies had a mean of 0 and standard deviation of 1.

For all keystrokes together, we predicted the variable z_{IKI} using the features below. We also grouped these reaction times into two bins. The first keystroke of a word, $z_{first-IKI}$, is the time required to type just the first letter of the current word. All other letters are "word-internal" keystrokes, $z_{internal-IKI}$. As described in the introduction —and supported by inner/outer loop models of language production —it may be that word-level frequency statistics only influence word access (e.g., first letter $z_{first-IKI}$). However, an effect of word-level statistics on $z_{internal-IKI}$ would suggest high-level word access and low-level motor control are dynamically coupled supporting cascade-based accounts of language production.

There are obvious factors that guide interkey timing, as described in the introduction to this paper. We used a number of predictor variables in a mixed effects regression model to capture latencies. For convenience, a representation of the final source of data for our regression models is shown in 5.2.

5.2.5 Control predictors

Digraph covariate. Our primary control variable was the identity of the so-called "digraph" of the interkey event. The time to type the letter 'e,' for example, will be determined quite significantly by the biomechanical constraints of the finger movement *from the prior character*.

There are several different ways to control for biomechanics-based variation in interkey timing including considering the IKIs for *each* bi-letter sequence, taking the intercept (or average) and considering this for each two letter sequence (i.e., each digraph). However, there are known effects of bi-letter frequency, tri-letter frequency and even one letter *after* the current letter (Gentner et al., 1988) suggesting the dynamics of one's IKIs do not occur in isolation. There is a frequency-based effect that is captured by models built solely to control for the biomechanics of one's hand placement. To avoid overfitting the digraph model, as we expect previous studies have done accidentally, we define the digraph by classifying different bi-letter combinations into categories based on finger placements. As a result the digraph consists of

[4]There is ongoing debate about how to specify random effects. Should we maximize the random-effect structure (Barr, Levy, Scheepers, & Tily, 2013)? Or, should we use PCA on the variance-covariance structure to estimate model complexity (Bates, Kliegl, Vasishth, & Baayen, 2015)? We take the more complicated, but conservative approach above, but also replicate our results, as described below.

Table 5.2: Example section of keystroke data used in analyses

$c_{t'-1}$	$c_{t'}$	z_{IKI}	first key?	digraph $\bar{z}_{IKI}$	$f(c_{t'})$	I	I_{prior}	I_{next}	w_{t-1}	w_t	w_{t+1}
‘ ’	i	1.318	TRUE	0.490	6,871	4.091	2.923	0.538	him	i	think
‘ ’	t	-0.104	TRUE	0.269	9,739	8.920	3.910	4.836	i	think	it
t	h	-0.458	FALSE	-0.550	7,957	8.920	3.910	4.836	i	think	it
h	i	1.080	FALSE	-0.149	6,871	8.920	3.910	4.836	i	think	it
i	n	0.959	FALSE	-0.131	7,073	8.920	3.910	4.836	i	think	it
n	k	0.843	FALSE	-0.017	766	8.920	3.910	4.836	i	think	it
‘ ’	i	1.318	TRUE	0.490	6,871	5.009	2.122	2.407	think	it	is
i	t	2.025	FALSE	0.026	9,739	5.009	2.122	2.407	think	it	is
‘ ’	i	-0.103	TRUE	0.490	6,871	5.364	2.651	2.893	it	is	nice
i	s	-0.694	FALSE	-0.440	6,008	5.364	2.651	2.893	it	is	nice
‘ ’	n	2.735	TRUE	0.612	7,073	8.206	4.865	8.115	is	nice	how
n	i	0.606	FALSE	-0.044	6,871	8.206	4.865	8.115	is	nice	how

four categories of current and just-typed key corresponding to the fingers which typed them —same letter, same finger, different finger same hand, different finger different hand. We take the average of each individual's digraph class as the estimate of that individual's biomechanical movement. We expect that the average interkey time per category would, by itself, still account for much of our observed variance in character reaction times, while increasing the likelihood of accurately modeling frequency effects.

Character frequency. We predict there to be a negative relationship between character frequency, $f(c_{t'})$, and keystroke latency. This is a well known effect and stands as a second control predictor incorporated into our baseline model.

5.2.6 Predictors of Interest

Current-word information. We expect the probability of a word will significantly predict interkey times (Cohen Priva, 2010). We extracted bigram frequency information from the Corpus of Contemporary American English (COCA) in order to assess word and bigram probability (Davies, 2009). The probability of any word was assessed as the probability with which that word occurred in our bigram set. As expected, this distribution is quite long-tailed. To adjust, we calculated the negative log-probability as the covariate. We call this the current word's information value:

$$I = -log(p(w_t)) \tag{5.2}$$

This uses base-2 *log*, and so we will refer to it on a scale of *bits* of information.[5]

Prior- and Next-word information. As described in the introduction, the question at hand is whether interkey timing can be influenced by word context —namely, whether the prior and next word can speed up (or slow down) the typing of a current word. To test this, we look to information-theoretic measures "prior" and "next" information. Given a target word that is being typed at time t, the prior information is defined as the negative *log* probability of the current word given the prior word:

$$I_{prior} = -log(p(w_t|w_{t-1})) \tag{5.3}$$

The "next" information is defined in the same way, but uses the probability given the next word:

[5]The number of binary digits or *bits* of a word can be thought of as the number of yes/no decisions needed to accurately determine the probability of that word among all other words in the corpus.

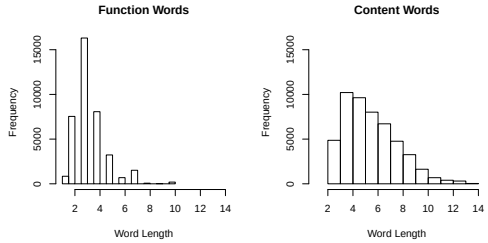


Figure 5.1: Function & Content word frequency by Length

$$I_{next} = -log(p(w_t|w_{t+1})) \tag{5.4}$$

Similarly, both of these predictors use base-2 *log*, and so we will refer to them on a scale of bits of information. An important caveat in this and any research using information theory is that these measures depend on the background probability distribution that is used. Here, we use COCA bigrams, and so a given number of bits reflects the information load in a word on the background of average American English usage.

5.2.7 Types of models

Function vs. Content words. We split our data into function and content words to determine whether the influence of current word information, prior word information or next word information more strongly predicts keystroke latencies in different parts of speech (Shih, 2014). For instance, function word IKIs may be influenced more by prior or next word information than current-word information because these words are more easily used to adjust information flow *in context*; across the entire message (Aylett & Turk, 2004; Jaeger, 2010). Alternatively, keystroke latencies in content words may be predicted by their current word information as the word itself portrays a level of contextual dependence regardless of its frequency among surrounding words ((e.g., prior or next word information).

Beyond the semantic/syntactic differences between content and function words, they also differ by frequency, inversely correlated with length. 5.1 shows the overall frequency of content and function words by length within our corpus. Summary statistics of the remaining models as well as their replication from data collected via Amazon Mechanical Turk can be found in the below.

Word length. Word length is often correlated with various frequency-based statistics such as raw word frequency (Zipf, 1949). This is a problem when trying to decipher what aspects of the language influence its production: spurious correlations among predictor variables can result in inflated, or flipped beta values (Wurm & Fisicaro, 2014). Since we are interested in how frequency-based statistics of a word impact production dynamics beyond word length, and because our dataset consists of hundreds of thousands of data points, we ran regression models split by word length to avoid problematic multicollinearity. These models happen to be different for function and content words, due to the nature of their distributions (e.g., seven-letter function words almost never occur). We report on the most frequently typed word length for content words ($length = 4$), and function words of length three through six. As evidence that our effects are not simply due to our selection of specific word lengths (however justified this might be) we replicate the effects of content words in a full linear mixed effects model, which can be found below.

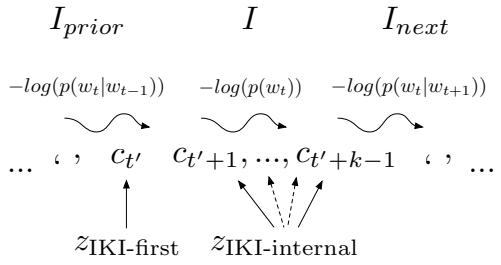


Figure 5.2: Effects of Word Level Information on Inter-Key-Intervals

5.2.8 Predictions

As we discussed in the introduction, one hypothesis about the influence of word level information is that they have distinct facilitative effects (see Figure 5.2). We expect all these probabilistic lexical variables contribute to all interkey interval types (all, first, internal). However, it may be the case that these variables have distinctive effects. First, the current word information of a given word will directly predict both first and internal key times:

$$z_{first-IKI}, z_{internal-IKI} \propto I \tag{5.5}$$

We predict prior-word information will dictate the latency of the initial keystroke on the current word, because the prior word facilitates access to the current word. Words less predictable by the prior word —and thus have higher information load associated with them —are more likely to have a slower initial keystroke. In other words, the first key is proportional to prior information:

$$z_{first-IKI} \propto I_{prior} \tag{5.6}$$

What about the contribution of next information? We anticipate this probability facilitates initial keystroke, too, as we expect this marks the word-access gains from predictable words. However, another possibility is that the word “flows” through interkey timing and keystrokes within a word. If so, the final keystroke of a word, will be “pulled” by the subsequent word ($t+1$).

$$z_{internal-IKI} \propto I_{next} \tag{5.7}$$

A key question the analyses below answer is whether these relationships differ in function vs. content words. Intuitively, these word classes have distinct roles in structuring text. Function words are highly frequent, and frame content words, which have a much higher type-token ratio (Monaghan et al., 2005). So while separating part-of-speech analyses gives us a basis for controlling for frequency and length, interkey times may also reveal cognitive dynamics that mark these linguistic roles. After sharing the results, we revisit implications of this in the General Discussion.

5.2.9 Participant descriptives and data processing

Because several participants used browsers or keyboards not compatible with our JavaScript code, or because some participants did not provide adequate text for analysis (empty submissions), only 265 participants (84%) had data appropriate for analysis. The 52 who did not were discarded and not included in any analysis.

The participants submitted an average of 149.7 words (827.2 keystrokes), with a maximum contribution of 1,067 words (5,799 keystrokes) from one participant, and a minimum of 9 words (43 keystrokes). In total, we collected 219,206 interkey times from our 265 participants.

We did not include all keystrokes in our analysis. First, we treated any latency of 500ms or more as a hesitation, following recent typing research (Chukharev-Hudilainen, 2014; Cohen Priva, 2010). Excluding these times resulted in 88.2% (n=193,491) keystrokes remaining in our analysis.

We also excluded space and period characters, since these are not part of the words over which we have bigram or frequency information. Excluding these two common characters reduces our keystroke set to 153,986. Though, in principle, it would be valuable to analyze spaces and period characters in our set, we used the space character to delimit tokens, to which we associated lexical frequency and bigram frequency information. For this reason we focus instead on alphanumeric characters. Note, as seen in Table 5.2, that we did allow these characters to be the first ones, $c_{t'-1}$, for the digraph variable.

We only included keystrokes that were within words for which we had frequency information. Of course, participants often misspell words or use nonce or slang terms that are not in the COCA set. Despite this fact, out of the 153,986 keystrokes remaining, we had a set of times 98,410 (64%) remaining. This became the core set of interkey intervals which we modeled using mixed linear regression.[6]

5.2.10 Model Analysis

We built multiple linear mixed effects regression models with the outcome variable subject-based z-scores, z_{IKI}, predicted by digraph latency (average z-score for the given digraph class by participant), current letter frequency, prior information, log of word frequency, and next information for *each* word length by function and content words. All predictor variables were centered and scaled prior to analysis.

We compare baseline models containing both random effects (current word, *cw*, and subject, *s*) and our control predictors using the *lme4* package in *R*: $lmer(z_{IKI} \sim digraph + f(c) + (cw|1) + (s|1))$ —against models containing predictor variables of interest by comparing their Bayesian Information Criterion values ($BIC_{BL-Full}$). Raftery (1995, Table 6) provides a conservative reference to measure the relative difference in model fit. Larger positive differences indicate the observed data is more likely to occur given the full model. For example, $BIC_{BL-Full} > 10$ indicates the odds that the observed data occurring under the assumptions we make in our model (e.g., knowledge about word-level frequency values) compared to a null model are around $150:1$. This reflects a $> 99\%$ chance that our model is likely to be more accurate given our data. The difference in models can be used to calculate Bayes Factor which provides a direct measure of this posterior odds ratio. Bayes Factor is be calculated by taking the inverse of the $log_2(Bayes\ Factor) = \Delta(BIC_{BL-Full})$ which corresponds to $Bayes\ Factor = 2^{BIC_{BL-Full}}$. "Negative" Bayes Factor values, simply indicate odds ratios in favor of baseline models and are calculated taking the difference, $\Delta(BIC_{BL-Full})$. We report *BIC* values for the baseline model, the full model, and $BayesFactor$ for the comparison between the baseline and full models. A negative $BayesFactor$ value means the data more likely occurred under a model that does not assume additional knowledge of word level characteristics.

We also determined the amount of variance accounted for by each model using a pseudo R^2 measure (Nakagawa & Schielzeth, 2013): R^2_c corresponds to the amount of variance accounted for the full model, whereas R^2_m corresponds to the variance accounted for by fixed factors only. When comparing baseline and full models, subtracted baseline R^2_m values from the full model R^2_m values, $\Delta R^2 = \Delta R^2_{full} - \Delta R^2_{BL}$. This provides a measure of the amount of variance in our data that can be accounted for by our word-level predictors alone.

[6] It is also important to note at this juncture that we present two full replications of our study in below, where all our basic effects still obtain under these data processing procedures.

5.2.11 Results

Content words

Prior to analysis we first determined if our predictor variables were correlated. As with any naturalistic, observational corpus, several variables are inevitably correlated. Such multicollinearity can influence the interpretation of the contribution of individual predictors. To control for multicollinearity without altering the dataset significantly, we explored different length content words, starting with the most frequent, four-letter words ($n = 12,712$). The predictor variables for four-letter content words did not present problematic multicollinearity. Moreover, the observed effects from all word lengths show a stable effect of the same predictor variables found in four-letter word lengths. Therefore, the present model stands as a representative sample of our larger dataset.

There is no standardly accepted prescription for dealing with collinearity: that is, it can't be totally "solved" —only ameliorated—in complex datasets. We report the effects of the full model as well, but note that beta values on the full model cannot be trusted to be fully accurate as stronger correlations can result in odd effects, such as beta value sign flips. We tried building this out by adding content words of lengths three and five, but any additional length problematically increased collinearity.

We compared the baseline model with an intercept-only model; the result demonstrates that the control predictors of within-subject digraph and character frequency significantly add to the goodness-of-fit of the model (with an odds ratio well over $150:1$). We use the baseline model for all following model comparisons.

5.3 shows the results of the mixed effects baseline model, the full model, and subset models broken down by first key, and internal keys for four-letter content words. *BL Bayesian Information Criterion* and *BL* R^2_m provide the baseline measures for each of the corresponding models. Across all regression models, the individual's biomechanics significantly predicted keystroke latencies ($p \ll .0001$). Further, character frequency significantly predicted keystroke latencies ($p \ll .0001$); more frequent characters are typed faster.

For the full model there is evidence, given the difference in BICs between baseline and full models, that current word-level information is positively related to keystroke latencies. For four-letter content words, a model that includes current word information is 12 times more likely to account for observed variance in keystroke latencies than a model that only includes the digraph and character frequencies. Current word information significantly predicted keystroke latencies in the full model and word internal models such that when a more information-dense word occurred (a less frequent word) keystroke latencies were longer. The difference in R^2_m values provides a relative measure of the variance accounted for by adding word-level predictors. In the case of the full model, current word frequency accounts for .7 of the observed variance ($\Delta R^2 = .7$) in keystroke latencies beyond our control variables.

This model shows that the full set of interkey times is predicted significantly by current word information. To investigate whether initial or word-internal keystrokes bear the weight of this informativity effect, we isolated keystrokes that were either at the onset of a word, or internal to a word. This would be in accord with simple, serial inner/outer loop theories in typing expertise, which see the onset of word-internal keystrokes as ballistic expert sequences that are uninfluenced by word level frequency statistics (Logan & Crump, 2011). We find that current word information better accounts for word-internal keystrokes, $z_{IKI-internal}$ ($p \ll .0001$) than for first key latency, $z_{IKI-first}$. Yet, the likelihood that the observed data occurred given our full model was well over $150:1$ for first key. From this we infer only that an individual's biomechanics or simply character frequency significantly account for the latency observed at the onset of a word, but that some knowledge of word-level frequency statistics does help improve the model's overall fit to the data.

The current word information alone contributes to our ability to predict interkey latencies beyond baseline ($\Delta R^2 = .011$). Crucially, the likelihood ratio for $z_{IKI-internal}$ provides weak

Table 5.3:
Multiple mixed effects linear regression models on 4-letter content words

	Dependent variable:			
	BL z_{IKI}	z_{IKI}	$z_{IKI-first}$	$z_{IKI-internal}$
	(1)	(2)	(3)	(4)
$\bar{z}_{IKI,digraph}$	0.357^{***}	0.356^{***}	0.392^{***}	0.251^{***}
	(0.009)	(0.009)	(0.068)	(0.008)
$f(c_{i'})$	-0.160^{***}	-0.157^{***}	-0.136^{**}	-0.089^{***}
	(0.008)	(0.008)	(0.033)	(0.008)
I_{prior}		0.007	0.010	0.008
		(0.005)	(0.014)	(0.005)
I		0.130^{***}	0.166	0.128^{***}
		(0.026)	(0.070)	(0.026)
I_{next}		-0.002	0.034	-0.013
		(0.015)	(0.045)	(0.014)
Constant	0.103^{***}	-0.008	0.359	-0.117^{*}
	(0.014)	(0.035)	(0.112)	(0.034)
Observations	12,712	12,712	2,505	10,207
BL Bayesian Inf. Crit		34,416.150	8,306.322	24,257.018
Bayesian Inf. Crit.	34,416.150	34,412.580	8,323.845	24,256.580
BL R^2_m		0.152	0.027	0.099
R^2_m	0.152	0.159	0.039	0.110
Bayes Factor		11.87	$\gg$150	1.35
Note:				*p<0.001; **p<1e-04; ***p<1e-05

(though significant) evidence that our data support a model containing current word information. To be clear, the Bayes Factor itself takes into account the complexity added to the model when it contains more fixed factors. When modeling $z_{IKI-internal}$ the addition of three new fixed factors (e.g., word level frequency factors) contributes to modeling the observed data (slightly) beyond the additional complexity they add to the model. It is likely the case that the effect of word frequency is strong enough to mitigate additional complexity added by more than one new fixed effect. With this in mind, the lexical influences on interkey timing are not reserved exclusively to the onset key of a word. They extend into the typing of the other letters in the word too. This is further elucidated by Figure 5.3 where we plotted the IKIs for content words broken down by length.

Table 5.4:
Multiple mixed effects linear regression models on all content words

	Dependent variable:			
	BL z_{IKI}	z_{IKI}	$z_{IKI-first}$	$z_{IKI-internal}$
	(1)	(2)	(3)	(4)
$\bar{z}_{IKI,digraph}$	0.328***	0.328***	0.388***	0.250***
	(0.004)	(0.004)	(0.046)	(0.004)
$f(c_{i'})$	−0.156***	−0.156***	−0.111***	−0.106***
	(0.004)	(0.004)	(0.018)	(0.004)
I_{prior}		0.002	0.019	0.003
		(0.002)	(0.007)	(0.002)
I		0.072***	0.080	0.093***
		(0.011)	(0.034)	(0.010)
I_{next}		0.003	−0.005	0.001
		(0.007)	(0.022)	(0.007)
Constant	0.099***	0.012	0.330***	−0.079***
	(0.006)	(0.016)	(0.064)	(0.017)
Observations	59,924	59,924	9,417	50,507
BL Bayesian Inf. Crit		161,821.100	30,673.69	124,846.5
Bayesian Inf. Crit.	161,821.100	161,763.000	30,684.010	124,730.400
BL R^2_m		0.132	0.025	0.093
R^2_m	0.135	0.135	0.033	0.102
Bayes Factor		$\gg$100	$\ll$-100	$\gg$100
Note:				*p<0.001; **p<1e-04; ***p<1e-05

Figure 5.3 shows how information flows within words broken by high and low information-dense words. IKIs are longest when first starting to type a new word, but remaining letters are not typed at a uniform rate. Specifically, there is no clear bottoming out effect that we might otherwise expect from theories that treat word-internal IKI as shielded from word-level effects (Logan & Crump, 2011). Instead, we see a dynamic flow that is influenced by word-level information density.

When modeling all content words (5.4, the odds of our data occurring without making

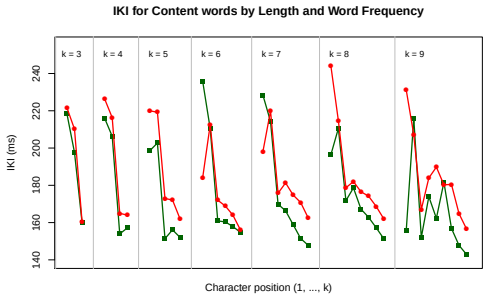


Figure 5.3: Content words separated by Length in characters (k)

Means are based on a median split on I. Green lines reflect more predictable words. This shows facilitated keystroke dynamics. In other words, when *precisely* the same sequence of keys is depressed predictability of that content word is reflected in the millisecond-level keystroke speed.

assumptions about word-level frequency statistics is well over $\gg 150:1$ for $z_{IKI-first}$. Even when modeling word-level statistics increases the amount of variance accounted for $\Delta R^2 = .008$, model fit is not substantially improved. However, the odds of our data occurring under word-level statistics assumptions is well over $\gg 150:1$ for $z_{IKI-internal}$.

Correlations among predictor variables were very high for the full content word model, in some cases $r > .75$. In the full model $z_{IKI-first}$ was not a substantially better fit than our baseline model. Indeed no other effects beyond digraph and character frequency were significant.

Function words.

Within function words, relatively high correlations between current word information and prior word information ($r = .77$) as well as current word information and next word information ($r = .74$) were observed. To mitigate these correlations, we first trimmed function words (orthographic letters) by length ($1 < words < 7$). This was done to control for collinearity; reducing all predictor variables to nonproblematic levels ($r < .69$).[7] The results are shown in Table 5.5.

We found a significant effect of digraph and character frequency as expected (BL z_{IKI}) capturing roughly 13% of the variance in keystroke latencies and roughly 2.2% beyond a simple intercept model ($R^2_c = .108$). The strength of the digraph covariate alone is to be expected from past research (Cooper, 1982). The full model also reveals that prior information contributes to capturing roughly .4% latency variance ($\Delta R^2 = .004$) beyond our baseline model. Prior information has a positive relationship to latency —words that encode more information, are typed more slowly ($p \ll .0001$). This effect holds when the data is divided by first key and internal keys. Prior word information captures roughly 1.7% of the variance in first keystroke latencies and 1% for internal keystrokes remaining, significant across all models.

A model that includes prior word information likely accounts for the observed keystroke latency data than a baseline model with odds $\gg 100:1$. Further, the amount of variance captured by prior word information alone may be accounting for some amount of additional complexity

[7] Another possibility is that clustering in the frequency of words encourages correlations. That is, perhaps there are two inherently different frequency distributions of function words due to the high use of only a few words. Such a skewed distribution may lead to inflated correlations where a more accurate model might treat them as separate groups. A prior analysis revealed that removing the most frequent term "the" from the dataset reduced collinearity substantially ($r < .60$). However, the results were not significantly different from those presented here.

Table 5.5:
Multiple mixed effects linear regression models on function words length 2-6

	Dependent variable:			
	BL z_{IKI}	z_{IKI}	$z_{IKI-first}$	$z_{IKI-internal}$
	(1)	(2)	(3)	(4)
$\bar{z}_{IKI,digraph}$	0.340^{***}	0.341^{***}	0.189^{***}	0.187^{***}
	(0.005)	(0.005)	(0.041)	(0.007)
$f(c_{i'})$	-0.051^{***}	-0.051^{***}	-0.143^{**}	-0.035^{***}
	(0.006)	(0.006)	(0.035)	(0.006)
I_{prior}		0.030^{***}	0.063^{***}	0.022^{***}
		(0.004)	(0.008)	(0.003)
I		-0.059	-0.010	-0.001
		(0.039)	(0.073)	(0.039)
I_{next}		0.023	0.040	0.015
		(0.012)	(0.027)	(0.011)
Constant	-0.006	-0.179^{***}	0.066	-0.229^{***}
	(0.023)	(0.035)	(0.079)	(0.035)
Observations	35,832	35,832	10,120	25,712
BL Bayesian Inf. Crit		92,274.340	30,694.85	56,889.37
Bayesian Inf. Crit.	92,274.340	92,251.240	30,672.710	56,894.920
BL R^2_m		0.130	0.015	0.052
R^2_m	0.130	0.134	0.032	0.062
Bayes Factor		$\gg$150	$\gg$150	-46.70
Note:				*p<0.001; **p<1e-04; ***p<1e-05

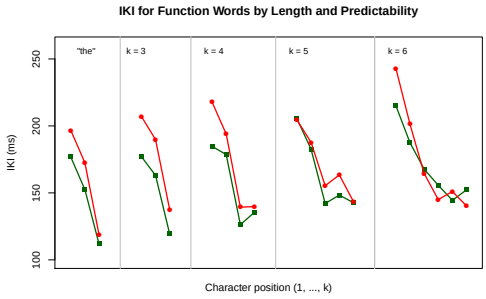


Figure 5.4: Function words separated by Length in characters (k)

Means are based on a median split on I_{prior}). Green lines reflect more predictable words given the prior word. These show facilitated keystroke dynamics, even in the more frequent and shorter words, and even in the most frequent function word, *the*. In other words, when *precisely* the same sequence of keys is depressed in this frequent function word, the predictability of that function word is reflected in the millisecond-level keystroke speed.

added by fixed effects factors (e.g., current word information and next word information). This highlights the strength of this effect. However, there is strong evidence that the observed data for internal keystrokes was more likely to occur under a more naive assumption, one that does not account for word level frequency effects. The negative Bayes Factor value for $z_{IKI-internal}$ indicates the posterior odds ratio is in favor of the baseline model (47:1). This simply means that while $z_{IKI-internal}$ *does* show a significant effect of prior word information, the addition of three new fixed factors (e.g., word level frequency factors) did not contribute enough to modeling the observed variance beyond the additional complexity they add to the model. In a follow up analysis we compared our baseline $z_{IKI-internal}$ to a model without current word frequency or next word information, which were not found to be reliable predictors of IKI latencies in the full model. We found strong evidence the observed keystroke latencies were more likely to occur given our knowledge of their prior word information only; $BayesFactor \gg 100$, $\Delta R^2 = .008$. In this new model all three fixed effects were significant. Figure 5.4 shows z_{IKI} by *length* split by high and low prior word information.

"The"

As an example case, we isolated the highest frequency word, "the," and ran the same models. The reason for testing this specific word is to see whether the effect of prior word information influences words thought to be isolated from contextual effects, due to their high frequency and ubiquitous nature across contexts (e.g, the expert use). Table 5.6 shows the model results without current word information (due to rank deficiencies as word frequency does not vary for a single word). The digraph model, again, is a highly significant predictor of keystroke latencies ($p \ll .0001$). There was no effect of character frequency (most likely due to the lack of variation among only three letters). The intercept model (not present in Table 4)accounted for roughly 4% of the variance in keystroke latencies ($R^2_c = .038$) while the baseline model alone captured 10.3% ($R^2_m = .103$, $\Delta R^2 = .065$). In the full model, prior word information was highly significant as expected ($p \ll .0001$), accounting for an additional 2.2% of the variance from baseline ($\Delta R^2 = .022$) and accounted for a substantial amount of variance in first key latencies ($\Delta R^2 = 0.046$) and word internal keystroke latencies ($\Delta R^2 = .022$). In each model, the inclusion of word-level frequency statistics not only significantly contribute to the variance accounted for,

but substantially improves model fit.

Table 5.6:
Multiple mixed effects linear regression models on the word "The"

	Dependent variable:			
	BL z_{IKI}	z_{IKI}	$z_{IKI-first}$	$z_{IKI-internal}$
	(1)	(2)	(3)	(4)
$\bar{z}_{IKI,digraph}$	0.294^{***}	0.303^{***}	-0.068	-0.002
	(0.012)	(0.012)	(0.079)	(0.029)
$f(c_{i'})$	-0.019	-0.021		-0.003
	(0.016)	(0.016)		(0.013)
I_{prior}		0.130^{***}	0.256^{***}	0.097^{***}
		(0.011)	(0.030)	(0.010)
I_{next}		-0.015	0.020	-0.032
		(0.031)	(0.075)	(0.029)
Constant	-0.248^{***}	-0.554^{***}	-0.419	-0.844^{***}
	(0.024)	(0.056)	(0.143)	(0.060)
Observations	5,388	5,388	1,489	3,899
BL Bayesian Inf. Crit		12,924.250	4,449.854	7,674.52
Bayesian Inf. Crit.	12,924.250	12,824.260	4,399.699	7,633.288
BL R^2_m		0.103	<.001	<.001
R^2_m	0.103	0.125	0.046	0.022
Bayes Factor		$\gg$150	$\gg$150	$\gg$150
Note:				*p<0.001; **p<1e-04; ***p<1e-05

The observed keystroke latencies for the most frequently occurring word were well over 100 times more likely to have occurred given our knowledge of prior word information (Bayes Factor $\gg$ 100). When occurrences of the words "the" are less frequent given their local context (prior word) all three letters are more likely to be typed at a slower rate.

5.3 Follow-Up Replication

We replicated the experiments herein twice to ensure that the stability of the results. We summarize these in the below. We used Amazon Mechanical Turk in the same basic task, and an additional task again for a set of participants in the UC Merced subject pool. We collected a number of subjects estimated from a simulation power analysis from the original observed data (in particular, how many subjects we required to achieve a 95% chance of replicating our results with $p < .0001$). The results are virtually the same, excepting some noise to be expected from a smaller subset of data for each, and some intriguing appearance of the by-next word predictability (which did not obtain in the first experiment, reported here). Specifically, in these replications we found current word information still predicts content words, prior word information predicts function words, and both current and prior word information predict word internal keystroke latencies over and above first key latencies.

5.3.1 Replication 1: AMT

Keystroke latencies were collected from Amazon Mechanical Turk users via an online task in March, 2016. *N* after selecting out subjects without or missing data = 128. Total number of keystrokes: 27,169. This is significantly less than the original dataset. As such echoes cognitive effects may be less detectible.

Content words

Collinearity among the full set of content words Table 5.7 was high. However, it is only with this larger dataset that we might detect the subtle effects of word frequency, however confounded they may be. We replicated word internal keystroke findings that the frequency of the current word is proportional to word internal latencies $z_{IKI-interal}$, ($p \ll .0001$)

Table 5.7:
Multiple mixed effects linear regression models on four-letter content words: AMT

	Dependent variable:			
	BL z_{IKI}	z_{IKI}	$z_{IKI-first}$	$z_{IKI-internal}$
	(1)	(2)	(3)	(4)
$\bar{z}_{IKI,digraph}$	0.317^{***}	0.317^{***}	0.362^{***}	0.265^{***}
	(0.007)	(0.007)	(0.056)	(0.007)
$f(c_{i'})$	-0.119^{***}	-0.119^{***}	-0.079	-0.082^{***}
	(0.007)	(0.007)	(0.028)	(0.007)
I_{prior}		0.010	0.032	0.008
		(0.004)	(0.013)	(0.004)
I		0.065^{*}	0.020	0.101^{***}
		(0.019)	(0.058)	(0.018)
I_{next}		0.007	0.008	0.002
		(0.013)	(0.040)	(0.012)
Constant	0.003	-0.125^{***}	0.130	-0.206^{***}
	(0.009)	(0.028)	(0.096)	(0.028)
Observations	15,511	15,511	2,728	12,783
BL Bayesian Inf. Crit		41,282.650	8,875.06	30,932.13
Bayesian Inf. Crit.	41,282.650	41,266.840	8,901.472	30,879.910
BL R^2_m		0.126	0.033	0.114
R^2_m	0.126	0.132	0.040	0.127
Bayes Factor		$\gg$150	$\ll$-150	$\gg$150
Note:				*p<0.001; **p<1e-04; ***p<1e-05

Four-letter content words

Collinearity for four-letter content words Table 5.8 was low. However, the dataset was substantially smaller. No effects of word-level frequency were detected when modeling only four-letter content words

Table 5.8:
Multiple mixed effects linear regression models on content words: AMT

	Dependent variable:			
	BL z_{IKI}	z_{IKI}	$z_{IKI-first}$	$z_{IKI-internal}$
	(1)	(2)	(3)	(4)
$\bar{z}_{IKI,digraph}$	0.346^{***}	0.345^{***}	0.361^{**}	0.285^{***}
	(0.016)	(0.016)	(0.084)	(0.015)
$f(c_{i'})$	-0.111^{***}	-0.110^{***}	-0.045	-0.070^{***}
	(0.015)	(0.015)	(0.053)	(0.014)
I_{prior}		0.028	0.037	0.028
		(0.009)	(0.027)	(0.009)
I		0.044	-0.164	0.096
		(0.044)	(0.124)	(0.044)
I_{next}		-0.014	0.029	-0.030
		(0.028)	(0.081)	(0.028)
Constant	0.026	-0.180	0.168	-0.278^{***}
	(0.020)	(0.060)	(0.182)	(0.059)
Observations	3,508	3,508	750	2,758
BL Bayesian Inf. Crit		9,677.666	2,572.567	6.704.593
Bayesian Inf. Crit.	9,677.666	9,700.489	2,601.842	6,717.216
BL R^2_m		0.129	0.026	0.123
R^2_m	0.129	0.135	0.029	0.136
Bayes Factor		$\ll$-150	$\ll$-150	$\ll$-150
Note:			*p<0.001; **p<1e-04; ***p<1e-05	

Function words

There was substantial collinearity among function words between one- and seven-letters. Table 5.9 However we report their results here as a direct replication. We replicated the effect of I_{prior} prior word information such that the onset of less predictable function words given the word it comes after is likely to be slower.

Table 5.9:
Multiple mixed effects linear regression models on Function words (2-6): AMT

	Dependent variable:			
	BL z_{IKI}	z_{IKI}	$z_{IKI-first}$	$z_{IKI-internal}$
	(1)	(2)	(3)	(4)
$\bar{z}_{IKI,digraph}$	0.321^{***}	0.322^{***}	0.266^{***}	0.226^{***}
	(0.009)	(0.009)	(0.044)	(0.012)
$f(c_{i'})$	-0.052^{***}	-0.052^{***}	-0.096	-0.029
	(0.011)	(0.011)	(0.044)	(0.012)
I_{prior}		0.022^{*}	0.055^{**}	0.013
		(0.006)	(0.014)	(0.007)
I		-0.018	0.104	0.046
		(0.056)	(0.101)	(0.052)
I_{next}		0.007	0.067	-0.027
		(0.019)	(0.042)	(0.020)
Constant	-0.074	-0.192^{*}	0.038	-0.250^{***}
	(0.027)	(0.052)	(0.110)	(0.051)
Observations	10,806	10,806	3,326	7,480
BL Bayesian Inf. Crit		26,904.04	9,670.295	16,091.94
Bayesian Inf. Crit.	26,904.040	26,937.550	9,680.359	16,129.790
BL R^2_m		0.117	0.027	0.076
R^2_m	0.117	0.159	0.039	0.085
Bayes Factor		$\ll$-150	$\ll$-150	$\ll$-150
Note:				*p<0.001; **p<1e-04; ***p<1e-05

5.3.2 Replication 2: SONA

Keystroke latencies were collected from University students from the comfort of their own computers via an online task. This data was collected in April, 2016. One year after the original data were collected. *N* after selecting out subjects without or missing data = 180. Total number of keystrokes: 61,011. Again this ls less than the original dataset. As such echoes cognitive effects may be less detectible.

Content words

We replicate significant effects of word frequency for the full content model(s) above Table 5.10. This was significant for z_{IKI} and $z_{IKI-internal}$ but not for $z_{IKiI-first}$. Again, this supports our hypothesis that word frequency influences word internal dynamics, and is not driven by the effects they present (if any) on the latency fo the first character of a word. Interestingly, in this model, next word information was significant, suggesting that the the likelihood of the current word occurring before the next word is proportional to the time is takes to type that word. If true, would suggest current language production is influenced by our anticipation of the next word. This is speculative at best, and will require more data to further elucidate these effects.

We also note that keystroke latencies for all letters, and for word internal letters lend strong evidence in support of a model that assumes knowledge of word-level frequency effects (odds being greater than 150:1). However, when modeling only the first letter of each word, the model is not substantially improved by assuming knowledge of word level frequency. This is a direct replication of the full models for both previous university student data, and Amazon Mechanical Turk data.

Table 5.10:
Multiple mixed effects linear regression models on content words: SONA

	Dependent variable:			
	BL z_{IKI}	z_{IKI}	$z_{IKI-first}$	$z_{IKI-internal}$
	(1)	(2)	(3)	(4)
$\bar{z}_{IKI,digraph}$	0.340^{***}	0.340^{***}	0.362^{***}	0.262^{***}
	(0.005)	(0.005)	(0.050)	(0.005)
$f(c_{i'})$	-0.139^{***}	-0.139^{***}	-0.107^{***}	-0.087^{***}
	(0.005)	(0.005)	(0.023)	(0.005)
I_{prior}		0.004	0.028	0.003
		(0.003)	(0.010)	(0.003)
I		0.049^{*}	0.015	0.077^{***}
		(0.014)	(0.042)	(0.014)
I_{next}		0.028^{*}	0.039	0.023
		(0.009)	(0.027)	(0.008)
Constant	0.044^{***}	-0.051	0.246	-0.140^{***}
	(0.008)	(0.021)	(0.077)	(0.021)
Observations	34,714	34,714	5,846	28,868
BL Bayesian Inf. Crit		92,635.510	18,799.820	70,085.18
Bayesian Inf. Crit.	92,635.510	92,611.550	18,817.620	70,029.970
BL R^2_m		0.139	0.027	0.102
R^2_m	0.139	0.142	0.035	0.111
Bayes Factor		$\gg$150	$\ll$-150	$\gg$150
Note:				*p<0.001; **p<1e-04; ***p<1e-05

Four letter content words

We continue to show significant effects of word frequency Table 5.11, but in each case the data were more likely to occur given the baseline model than a model that assumes knowledge of three additional word-level statistics. The main effect of word frequency is again replicated. The discrepancy in fitness between the full and four-letter word models is most likely due to a reduction in sample size, such that making similar assumptions over a smaller dataset results in increased model complexity overshadowing the subtle effects of word frequency of keystroke latencies.

Table 5.11:
Multiple mixed effects linear regression models on 4-letter content words: SONA

	Dependent variable:			
	BL z_{IKI}	z_{IKI}	$z_{IKI-first}$	$z_{IKI-internal}$
	(1)	(2)	(3)	(4)
$\bar{z}_{IKI,digraph}$	0.356^{***}	0.354^{***}	0.421^{***}	0.258^{***}
	(0.011)	(0.011)	(0.075)	(0.011)
$f(c_{i'})$	-0.148^{***}	-0.146^{***}	-0.135	-0.084^{***}
	(0.010)	(0.010)	(0.045)	(0.009)
I_{prior}		0.004	0.015	0.005
		(0.006)	(0.018)	(0.006)
I		0.127^{**}	0.091	0.134^{**}
		(0.033)	(0.091)	(0.032)
I_{next}		0.019	0.053	0.004
		(0.019)	(0.053)	(0.018)
Constant	0.058^{*}	-0.040	0.258	-0.152^{*}
	(0.018)	(0.042)	(0.137)	(0.042)
Observations	8,051	8,051	1,607	6,444
BL Bayesian Inf. Crit		21,583.900	5,217.694	15,328.55
Bayesian Inf. Crit.	21,583.900	21,591.120	5,244.523	15,337.920
BL R^2_m		0.150	0.037	0.102
R^2_m	0.150	0.157	00.047	0.113
Bayes Factor		-149.735	$\ll$-150	$\ll$-150
Note:				*p<0.001; **p<1e-04; ***p<1e-05

Function words

Prior word frequency, again, is highly significant across all models including $Z_{IKI-internal}$ Table 5.12. This effect replicates our findings. Our data are more likely to have occurred assuming knowledge of word-level statistics when we aim to predict all keystroke latencies and also when we simply predict first keystroke latencies again replicating our main paper findings.

Table 5.12:
Multiple mixed effects linear regression models on function words: SONA

	Dependent variable:			
	BL z_{IKI}	z_{IKI}	$z_{IKI-first}$	$z_{IKI-internal}$
	(1)	(2)	(3)	(4)
$\bar{z}_{IKI,digraph}$	0.340^{***}	0.340^{***}	0.221^{***}	0.185^{***}
	(0.007)	(0.007)	(0.048)	(0.009)
$f(c_{i'})$	-0.047^{***}	-0.047^{***}	-0.147^{**}	-0.042^{***}
	(0.008)	(0.008)	(0.036)	(0.008)
I_{prior}		0.033^{***}	0.072^{***}	0.023^{***}
		(0.004)	(0.011)	(0.004)
I		-0.037	0.138	0.011
		(0.043)	(0.075)	(0.044)
I_{next}		0.015	-0.008	0.023
		(0.014)	(0.032)	(0.013)
Constant	-0.036	-0.213^{***}	0.058	-0.277^{***}
	(0.024)	(0.039)	(0.089)	(0.038)
Observations	24,416	24,416	6,987	17,429
BL Bayesian Inf. Crit		63,268.660	21,649.17	37,909.41
Bayesian Inf. Crit.	63,268.660	63,261.920	21,629.830	37,922.010
BL R^2_m		0.127	0.020	0.054
R^2_m	0.127	0.133	0.052	0.069
Bayes Factor		106.866	$\gg$150	$\ll$-150
Note:				*p<0.001; **p<1e-04; ***p<1e-05

5.4 General Discussion

A general and central implication of our results is that first keystroke alone did not relate to language statistics. Even for function words, the effects of surrounding words influence various aspects of word-internal key presses. We show that language production (at least in typing) is dynamic at the smallest measurable unit (individual letters). Our results provide evidence that typing is not driven solely by the serial inner loop/outer loop theories, which treat the timing of word-internal keystrokes as ballistic non-varying expert sequences (Logan & Crump, 2011). Lexical influences on interkey timing are not reserved exclusively to the onset key of a word. They extend into the typing of the rest of the word, too.

The example case of function word "the" shows the effects of prior word on even the most frequently typed word in English. In fact, function words appear to show a stronger relationship to multi-word statistics than content words. One reason for this could be the structuring role of function words in the statistics of natural language. Function words likely have a ceiling effect in their relatively high frequency. Yet, the specifics of the context, such as prior word, impact their relative predictability. The dynamic unfolding of function words might help serve as predictive frames for their surrounding contexts (Mintz et al., 2002; Valian & Coulson, 1988; Monaghan et al., 2005) ensuring levels of efficient communication (Aylett & Turk, 2004; Jaeger, 2013). The data we present here show that these statistical tendencies play out in real-time performance measures —how fast or slow language is being produced at the fingertips.

We defined our information values using a general English corpus, the Corpus of Contemporary American English, reflecting the population's actual distribution. This lends further support to our findings, that during unscripted typing, function words are influenced by prior word information while content words are not. One explanation, is that the very nature of function words —highly frequent across all forms of communication —makes any fluent speaker an expert function word user. This expertise perhaps constrains the distribution of their production to be more ballistic and less influenced by a word's statistical properties *in isolation*. Ironically, it is in this *expert* language use that we detect the presence of broader contextual effects, and the flow of information.

5.4.1 Information and Dynamics

Our results imply a role for probabilistic expectancies in fine-grained language performance. This is consistent with many prior and recent studies (Huettig, 2015; Van Petten & Luka, 2012; Shaffer, 1973; Aylett & Turk, 2006, 2004; Tily et al., 2009; Mahowald et al., 2013; Levy & Jaeger, 2006; Frank & Jaeger, 2008; Genzel & Charniak, 2002; Jaeger, 2013, 2010). Our findings are consistent with a theory of language production that sees it as an adaptive system seeking a kind of efficiency (Jaeger, 2013).

Our results extend Smith and Levy's (2013) findings that reading time is logarithmically related to word frequency, in a production task. Similarly we find production speed is related logarithmically to word frequency. More so than the speed at which we read words, our production dynamics are sensitive to the statistical properties of language even when we define the words we encounter. Our effects are significant in that production is inherently proactive, being that word-level frequency influences the dynamics of the actions it takes to convey them. Further, we show effects that go beyond the current word, suggesting that the local context in which words are nested influences action as well. Our sensitivity to distributional effects of language during production, especially prior word information, highlights the cognitive system's dynamic contextual dependence as a continuous flow of information during mental processing (Spivey & Dale, 2006; Spivey et al., 2005). Our results conflict with an account that describes language production as a ballistic sequence of events whose timing is captured primarily by references to one's biomechanics (Logan & Crump, 2011). We are sensitive to the distributional properties of the communicative system, and this is reflected in our ecologically valid motor actions of language production. Our findings support a more general approach to cognition that features prediction as a primary cognitive process (A. Clark, 2013).

5.4.2 Future directions

Muscle Memory? It is possible that the highly practiced typing from participants —their "muscle memory" —is simply showing this echo of word-level statistics, rather than a "leaking out" of word processing into the keystrokes themselves. This may be possible, but we would still argue that our results are interesting. The subtle changes in keystroke timing from word predictability seen in Figures 5.3 and 5.4 means that such muscle memory effects

are encoding multi-word statistics. Some classical models of typing and expertise would predict that at the onset of a word, the remaining keystrokes should all flatline at their baseline levels (the digraph and character frequency estimators). However this is not what we observed, which means that, in some manner, the cognitive system would have to be storing these multi-word statistics in a manner that they reveal evidence of it in the movement of the effectors.

Other Linguistic Effects? Differences in character position findings suggest that there may be nonlinearities in the keystroke times (this can be observed in Figures 3 and 4). These could reflect morpheme boundaries. Some recent research on morpheme boundaries shows effects in expert typing tasks (Gagné & Spalding, 2014). The methods we describe here, with an expanded corpus, may permit us to test the effects of these morpheme boundaries in natural timing.

In a broader sense, our methods and results point to a fruitful integration of dynamical and information-processing approaches to cognition. Crucial to this integration is that the very task in which we obtain our results is spontaneously generated language production.[8] In real time, the information structure of one's spontaneous language use, akin to their online thought processes, influences the rate at which that thought is expressed. In this sense, comprehension and production may be bound by similar processes.

5.5 Conclusion

Typing can serve as a window into cognitive processes underlying language planning and its connection to motor action (Rumelhart & Norman, 1982). Typing provides an array of rich reaction-time data during planning and execution, and thus provides a kind of "micro-structure" of language performance. We collected *spontaneous* and *natural* language production via a simple online task and show that online tools for data collection permit rapid generation of large and natural typing corpora. We combine these millisecond-level data with current methods for measuring word predictability from "big data" (Behmer & Crump, 2015) and find that language may rely on prediction as a core cognitive process (Huettig, 2015; Van Petten & Luka, 2012; Shaffer, 1973), as may all of the human brain itself (Friston, 2010). Specifically, we found interesting differences in how prediction is used for content words and for functions words. Content words are more influenced by their frequency across a more general corpus of American English, while function words are more influenced by their frequency given the words produced right before them. Importantly, these influences of predictability show up not only on latency of the first letter of a word, but they spread into the later key presses for the rest of the word as well. Our findings suggest language production is not encapsulated, but dynamic and contextually dependent and support a cascade-based account of language production (Kandel et al., 2011; Van Galen, 1991; Will et al., 2006; Roux et al., 2013; Scaltritti et al., 2016; Olive, 2014). Our results show that the dynamics of action —at the fingertips —carry subtle echoes of single- and multi-word statistics.

[8] We provide the entire set of code and data used in this manuscript (https://github.com/DaveVinson/Spontaneous-typing-task). We share our tools here for both open access and replicability and to encourage researchers to utilize online collection methods to generate large corpora rapidly.

Chapter 6

Decision contamination in the wild: Sequential dependencies in online review ratings

6.1 Introduction

Humans are surprisingly bad at rating the absolute magnitude of their internal cognitive states. Regardless of the task, judgments of the absolute magnitude of a stimulus, experience, or feeling, are inherently contaminated by relative information from the sequence of judgments prior to the current one. Although we tend to believe that our judgment reflects the absolute value of the current experience, a good deal of the judgment is in fact determined by the relative difference between the current experience and experiences from previous trials (Laming, 1984; Stewart, Brown, & Chater, 2005). This pattern is complicated by the fact that decisions are independently influenced by factors such as stimulus, response, and feedback (Donkin, Heathcote, & Brown, 2015).

These cognitive sequential dependencies (SDs) occur whenever behavior on a trial is influenced by behavior on preceding trials. Far from rare, SDs are ubiquitous in cognition, contaminating absolute judgments from low-level perception all the way up to high-level moral judgments. We see the effect of previous trials on latency, accuracy, the type of errors produced, and interpretation of ambiguous stimuli. SDs seem to affect all levels of the cognitive system, including motor control (Dixon, McAnsh, & Read, 2012) , spatial memory (Freyd & Finke, 1984), face perception (Hsu & Yang, 2013; Liberman, Fischer, & Whitney, 2014), selective attention (Kristjansson, 2006), decision making (Jesteadt, Luce, & Green, 1977), and language processing (Bock & Griffin, 2000).

SDs have primarily been studied in the laboratory or at least with well-controlled experimental stimuli. They are more difficult to study in real-world scenarios because of the very large number of trials that would be required to identify their effects. In this paper, we explore SDs in a real-world situation by mining two and large natural datasets of online review ratings from (1) Yelp Inc. and (2) Amazon Inc. We use these datasets to determine if current review ratings are contaminated by previous reported experiences. First, we review SD trends observed in standard laboratory tasks.

6.1.1 SDs in the Laboratory

Assimilation occurs whenever the judgment of stimulus at time point n moves closer on the measurement scale to the judgment of stimulus k steps behind, at *n–k*, than it otherwise

would have been. Contrast is the opposite effect, when the judgment of stimulus n moves further away on the measurement scale from the judgment of stimulus n-k. In this sense, assimilation can be thought of as an attracting force from the preceding stimulus, while contrast can be thought of as a repelling force (Zotov, Jones, & Mewhort, 2011).

Much of the early work on SDs was psychophysical in nature and involved rating unidimensional stimuli such as the loudness of a tone or length of a line (Garner, 1953; Holland & Lockhead, 1968). Identifying the absolute magnitude of these stimuli (e.g., line length) has been well studied: Errors when identifying stimulus n assimilate towards the stimulus on trial n-1[1]. Participants are more likely to estimate the magnitude of a stimulus as more similar to the preceding stimulus when identifying it. Oddly, categorization of the same stimuli shows the opposite effect—a contrast effect from the previous response. When stimuli are clustered into categories and the response is a category label (e.g., small, medium, large), stimulus n is more likely to be labeled as belonging to a category further away on the measurement scale than stimulus $n-1$ (Stewart, Brown, & Chater, 2002; Ward & Lockhead, 1971).

The contrast effect (repelling) of trial n-1 on the category rating of trial n is not limited to low-level perception, but is seen across levels of cognition. As a striking high-level demonstration, consider Olivola and Sagara's (2009) findings that participants will elect to risk more human lives compared to a less risky alternative (with an equal probability of saving the same number of lives), when the number of lives at risk is equal to the probability of the number of lives lost when randomly selecting an observed disaster. The choice is clearly in contrast with one's experiences. Participants are willing to risk more human lives than average when there are a larger number of smaller casualty events. And less likely to risk more human lives when a higher number of high casualty events occur. Further, the binary choice (risky vs. sure) decision highlights the importance that this cannot be attributed to a scale-interpretation effect (i.e., artifact). Their findings further emphasize how statistical properties of our environment are reflected in our cognitive system. Similar patterns of SDs have been seen in a variety of laboratory tasks designed to tap real-world scenarios, including brake initiation latencies in driving behavior (Doshi, Tran, Wilder, Mozer, & Trivedi, 2012), jury evidence interpretation (Furnham, 1986), and clinical assessments (Mumma & Wilson, 1995). In addition, SDs seem to be immune to practice—they are seen even in overlearned and expert behaviors.

At first glance, SDs appear to be an irrational bias in decision making (or perhaps in event memory), and have been traditionally viewed as the natural by-product of low-level brain dynamics such as residual neural activation. However, more recent theoretical perspectives suggest that SDs may be a rational property of any cognitive system. These accounts characterize SDs in terms of an individual's adaptation to the statistical regularities of a nonstationary environment with related stimulus bundles (Qian & Aslin, 2014; Wilder, Jones, & Mozer, 2009; Angela & Cohen, 2009).

Computational models that explain how SDs emerge from the decision-making process are now being developed, at least for low-level perceptual tasks (Mozer et al., 2010). These models have great promise in that they may be reversed and then applied to rating data to "decontaminate" the rating, essentially producing a more accurate estimation of the individual's absolute experience of a product or business by removing the pollution from the relative information. Our interest is to mine large review datasets such as Yelp and Amazon, guided by knowledge from laboratory studies, to look for these naturally occurring contaminations that may affect how products and businesses are currently rated by reviewers and can expect to be rated in the future. In the case of Yelp, future business demand is largely influenced by online reviews (Cantallops & Salvi, 2014; Mudambi & Schuff, 2010) affecting a business's revenue between 5-9% with this number increasing by 50% for businesses with more than 50 reviews (Luca, 2016). This has an obvious benefit to the service quality Yelp and Amazon aim to provide, as well as a more accurate assessment of the products and businesses in question.

[1] Interestingly, the same absolute judgment that assimilates to the most proximal past judgment contrasts from stimuli $n-2\ldots5$.

In both Yelp and Amazon, reviewers rate their experience with a product or business on a scale of 1 to 5 stars. Because both the rating and rating scale are most similar to categorization tasks studied in the laboratory (i.e., what is the best label to classify the exemplar, experience with the business, on a scale of 1-5 stars), our predictions are loosely drawn from SDs in categorization. In particular, we expect that within reviewers we will see a contrast effect from ratings across products and businesses: For example, an individual's rating will be artificially inflated if his/her previous rating were lower, and artificially deflated if it were higher. In this sense, our predictions of review ratings are a simple extension of both the perceptual work of Zotov et al. (2011), and the moral judgments of Olivola and Sagara (2009).

Natural datasets are wrought with noise. Yet, where they lack structure they make up for in sheer size. We do not anticipate that SDs will play such a substantial role in altering the usefulness of user or business ratings on its face. Instead we expect to find echoes of these cognitive principles in large datasets of naturally occurring behavior. We consider this work a guided exploration, in an effort to bridge laboratory findings with relevant and functional natural behavior (Jones, 2016). In a sense, this is somewhat analogous to studying behavioral patterns of birds in aviary experiments to extrapolate to their foraging patterns in the wild.

6.2 Method

We used two datasets of online reviews, a Yelp, Inc. business review dataset and an Amazon product review dataset (movies & TV series[2]). While business and product reviews are inherently different, both are similar in that users rate their experience with the product or business. We used the most recent version of the Yelp Inc. data set at the time of research ("Round Seven"), released as part of Yelp's Dataset Challenge[3]. The dataset we used consists of just over 2.2 million reviews spanning 12 years from 2004-2016, with ratings between one (negative) and five (positive) stars ($\mu = 3.76$), from approximately 552,000 reviewers on roughly 77,000 businesses. Reviews were provided from nine cities (Edinburgh, Montreal, Karlsruhe, Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison) across four different countries (United States, Canada, Scotland and Germany). The Yelp review data is organized in a data format referred to as JSON ("JavaScript Object Notation"), and each line consists of a single JSON entry, for a user, review, etc. For our current analysis, we extracted the Yelp user's unique identifier, their star rating, and the time stamp of that rating. We then ordered the data by reviewer and date for further analysis. Star ratings follow a J-shaped distribution with mostly four and five star ratings, a dip in two star ratings, and roughly and equal number of one and three star ratings (Hu, Zhang, & Pavlou, 2009). The number of reviews increased steadily over Yelp's lifetime, consistent with Moore's Law. Similarly, the Amazon product review dataset—consisting of just over 4.6 million reviews dating as far back as 1998 (and up to 2015) only four years after its inception in 1994—also shows a J-shaped distribution over star ratings ($\mu = 4.19$) and increasingly more reviews over time. The downloaded Amazon dataset is organized in a CSV file consisting of the user's unique identifier, the item's unique identifier, star rating and time stamp.

In both Yelp and Amazon datasets, reviews occasionally occur at the same time (i.e., have the same time stamp). That is, even after organizing the data there will be some inherent 'noise' in our analysis since reviews that consist of the same time stamp by reviewer are randomly organized. This sort of noise is natural within larger datasets. However, due to the size of each dataset if there is a true SD effect, our analyses should not be affected by this noise. Distributions for both Yelp and Amazon can be seen in Figure 6.1.

[2]We extracted this dataset from *https://snap.stanford.edu/data/web-Amazon.html movies & TV reviews*. Further information can be found on this website including a recent update where obtaining the data requires emailing Julian McAuley (*julian.mcauley@gmail.com*) to obtain a link.

[3]Further information on how to access the dataset for free as part of Yelp's dataset challenge can be found at *http://www.yelp.com/dataset_challenge*

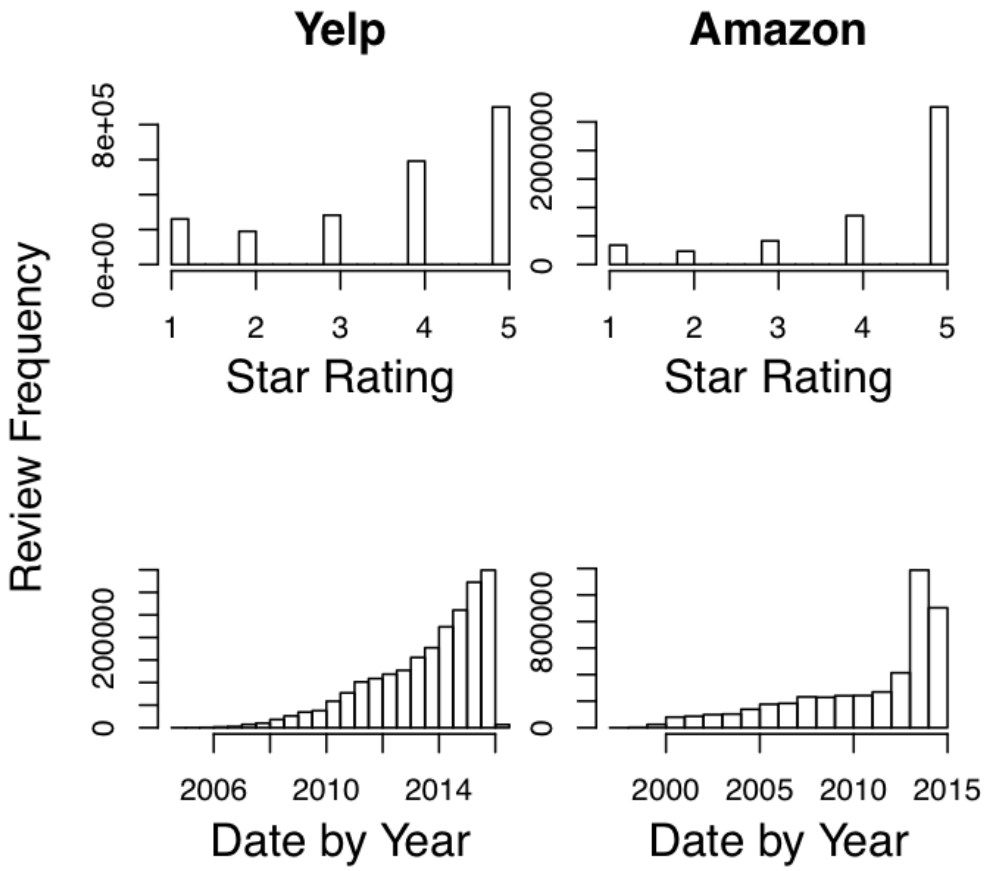


Figure 6.1: (top left) Frequency of Yelp reviews by Star Rating, (bottom left) frequency of Yelp reviews by year, (top right) Frequency of Amazon reviews by star rating, (bottom right) frequency of Amazon reviews by year

We tested whether previous review ratings influence the current rating within a user. If an individual's current review is sequentially dependent on the previous review, it will likely be the case that it will be repelled from previous reviews showing a contrast effect (Zotov et al., 2011). We anticipate that these effects will dissipate the farther away the previous review is from the current review. One possibility, and something we later address in the discussion section is the development of a measure of "bias" or "deviation" for each product / service by comparing a given review received by a given product to the average review received by the same product that was preceded by a median review (e.g., 4-star reviews).

6.2.1 Measures

We first calculated the deviation of the current review rating from its mean:

$$R_x - M(R_{T-x}) \tag{6.1}$$

Where R_x is the current rating and $M(R_{T-x})$ is the average rating for the reviewer with the current value x removed. The score is thus deviation of the current rating from the average of the prior ratings. This means that for a given user, the mean is dynamically adjusted over time. We used a mean adjustment, without standardization, because it can be interpreted in the original scale of the star ratings. In this way, for example, we can interpret +0.3 as 30% of the way to an entire increment of one "star" value. After obtaining this deviation score, we centered the values. It is important to note that the results are unlikely to be influenced as the standardization into z scores is likely only going to linearly translate the values, and leave the statistical patterns unchanged. This allows us to determine directly whether the reviewer's current rating is systematically biased away from his/her average response relative to the value of the preceding *n–k* review(s). To assess how distance is related to this deviation measure, we use *review distance* (*k*), an ordinal lag measure of the number of reviews (*k*) between the current review and previous review.

6.3 Results

6.3.1 Yelp

We first determined whether one's current review was related to his/her previous review rating at *k*-distances for Yelp reviews[4]. Figure 6.2 presents the mean and standard error bars for deviation of the current review rating from the mean (*y*-axis) by the previous star ratings (x-axis) at seven different review distances (*k*) within Yelp reviewers. The figure reveals an asymmetric contrast effect that dissipates the farther away the previous review is from the current review. At *n*–1 (the immediately preceding review) for example, a 1-star rating resulted in an increase in the subsequent rating from the overall mean rating. The opposite would be the case if the *n*–1 rating was 5 stars—the subsequent rating deviates toward a lower star rating relative to the average prior review ratings. In this sense, the data are very much consistent with Olivola and Sagara's experiment in that the current rating is systematically biased in the opposite direction from previous experience.

To assess these results quantitatively, we used eight linear models to predict current review ratings by *n–k* ratings for each of seven different values of *k* and a random review baseline. To create this baseline, we treated each value of *k* as distinct, thereby shuffling the results within reviewers. Besides reasons of computational and interpretive simplicity of linear regression, there are two additional reasons we employ this statistical method. First, the size of our

[4]All code used to visualize and analyze our results can be found here: $https://github.com/DaveVinson/sequential-dependence-reviews$

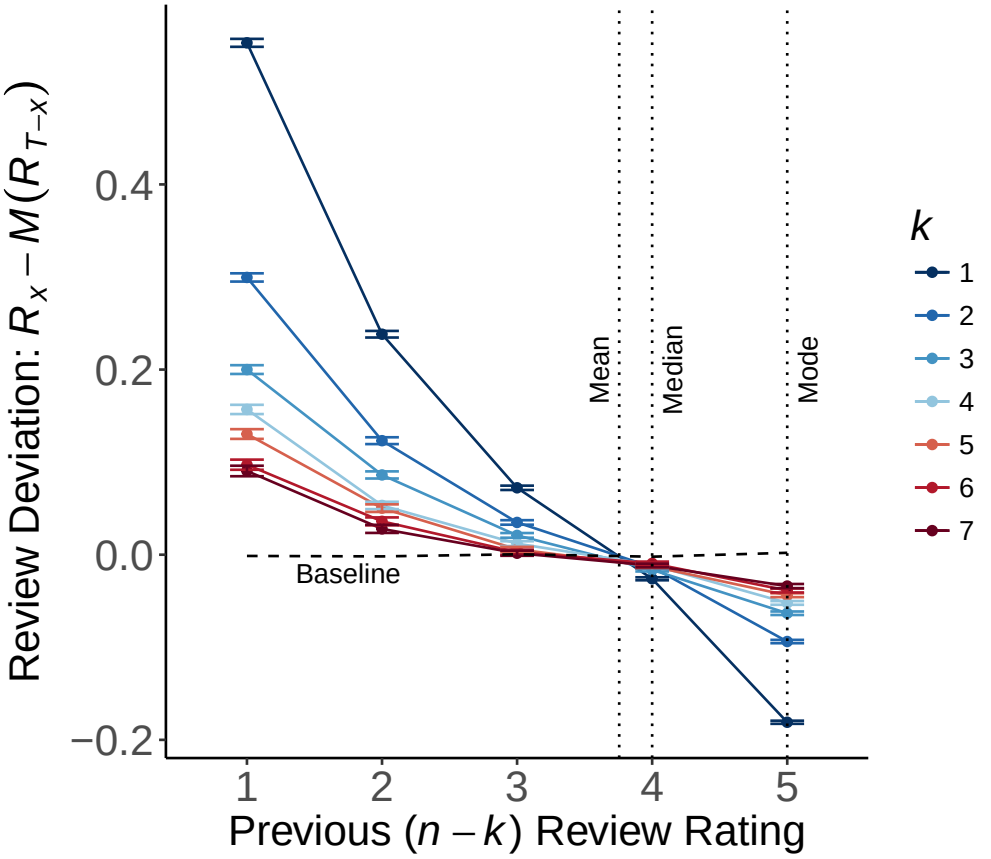


Figure 6.2: Deviation of the current review rating from the reviewer's average rating (y-axis) in relation to their previous review rating (x-axis) at k Review Distances for Amazon reviews

Table 6.1: Regression model for k Distances by Yelp Reviewer

k	BL 99.9%(CI)	F	$(d.f.)$	R^2_{adj}
Baseline	(−.001,.003)	.40	(1,1.9$x10^6$)	$<$ .0001
1	(−.17,−.16)	4.5×10^4	(1, 1.7×10^6)	.03
2	(−.09, −.08)	1.1×10^4	(1, 1.4×10^6)	.008
3	(−.06, −.06)	4,376	(1, 1.4×10^6)	.004
4	(−.05, −.04)	2,282	(1, 1.1×10^6)	.002
5	(−.04, −.03)	1,477	(1, 1.0×10^6)	.001
6	(−.03, −.03)	829	(1, 9.4×10^5)	.001
7	(−.03, −.02)	610	(1, 8.7×10^5)	$<$ .001

Note:CI is 99.9% confidence interval and d.f., the residual degrees of freedom equal to the number of observations for each k value.

dataset is quite large, including many thousands of individual reviewers; it is unlikely that non-independence of some observations will impact results. Second, we have straightforward linear hypotheses about the observed contrast effect, seen in Figure 6.2. In this way we are able to assess the relative linear impact each value of k has on the current review rating. We calculated this value in R using $lm([R_x - M(R_{T-x})]\ k)$. The results, presented in Table 6.1, reveal that as the value of k increases, as the current review is farther displaced from the previous review, the contrast effect dissipates. A randomly resampled review baseline where all reviews were first shuffled and then used to predict the reviewer's current rating ("Baseline" in Table 6.1) showed no significant effect on the current review rating. With the exception of the random review baseline, all values of k show a significant negative relationship with current review ratings, accounting for 2% of variance at the closest Review Distance ($k = 1$). The residual degrees of freedom in the F column of the table is equal to the number of observations in that category. At $k = 1$, the number of observations was 2.2×10^6 = 2.2 million reviews.

To determine whether there was a systematic decay in the effect of previous review ratings at different distances of k, we first subtracted the review value at each k distance from the current estimated review rating and then squared this value:

$$(R_x - M(R_{T-x}))^2. \tag{6.2}$$

We treat k as a continuous variable and use it to predict $(R_x - M(R_{T-x}))^2$. There was a significant negative relationship between k and the magnitude of its effect on the current review, $F(1, 8.2 \times 10^6) = 2.7 \times 10^4$, $R^2 = .003$, $CI = (-.067, -.064)$, $p < .001$ such that as k increases, the magnitude of the effect of the previous review decreases Figure 6.3.

6.3.2 Amazon

The contrast effect found within-reviewer Yelp ratings was replicated in the Amazon review ratings Figure 6.4. Like Yelp, the contrast effect dissipates the farther away the previous review is from the current review. The results, presented in Table 6.2, reveal that as the value of k increases the contrast effect dissipates. With the exception of the random review baseline, all values of k show a significant negative relationship with current review ratings, accounting for 1.5% of variance at the closest Review Distance ($k = 1$). At $k = 1$, the number of observations was 4.6$x10^6$; 4.6million reviews.

Again, we treat k as a continuous variable and predicted the magnitude of the observed contrast effect. There was a significant negative relationship between k and the magnitude of its

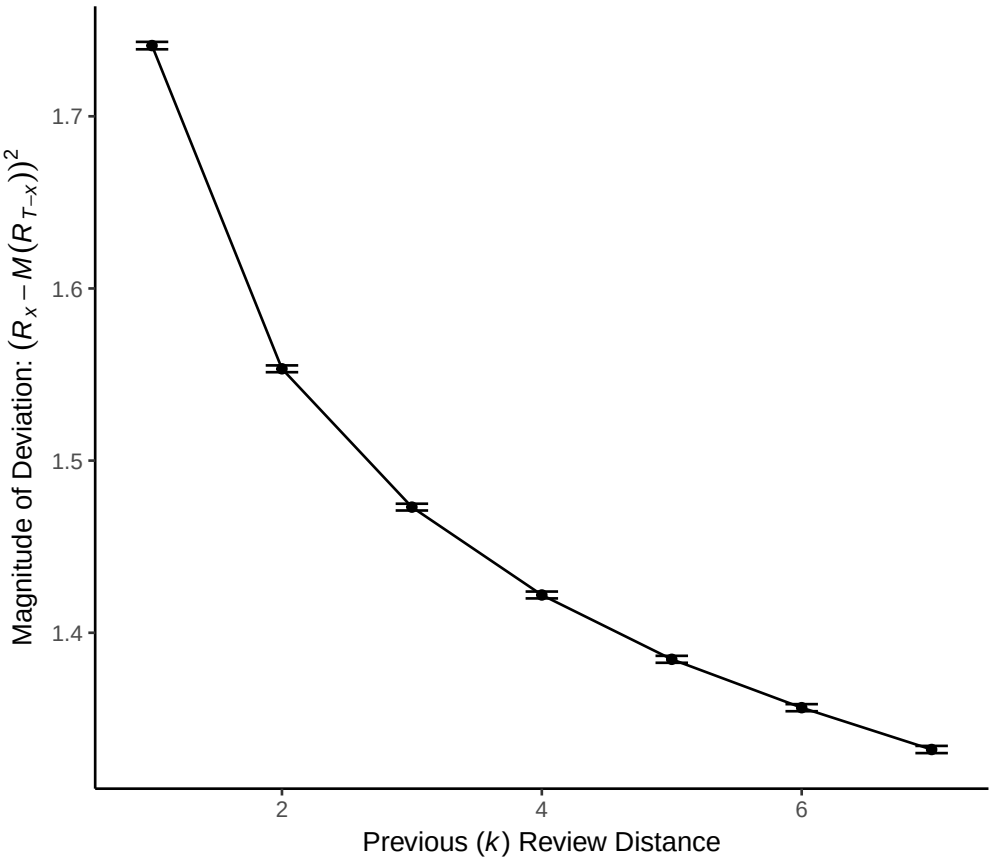


Figure 6.3: Magnitude of contrast effect on the current review at k distance from Yelp reviews

Table 6.2: Regression model for k Distances by Amazon Reviewer

k	BL 99.9%(CI)	F	$(d.f.)$	R^2_{adj}
Baseline	$(-.001,.002)$	1.36	$(1, 3.1 \times 10^6)$	$< .00001$
1	$(-.16,-.16)$	6.8×10^4	$(1, 2.5 \times 10^6)$	.026
2	$(-.07, -.07)$	1.1×10^4	$(1, 1.9 \times 10^6)$	.006
3	$(-.04, -.04)$	3,419	$(1, 1.6 \times 10^6)$	.002
4	$(-.03, -.03)$	1,423	$(1, 1.4 \times 10^6)$	.001
5	$(-.02, -.02)$	573	$(1, 1.2 \times 10^6)$	.001
6	$(-.02, -.01)$	277	$(1, 1.1 \times 10^6)$	.001
7	$(-.01, -.01)$	$1-5$	$(1, 1.1 \times 10^6)$	$< .001$

Note:CI is 99.9% confidence interval and d.f., the residual degrees of freedom equal to the number of observations for each k value.

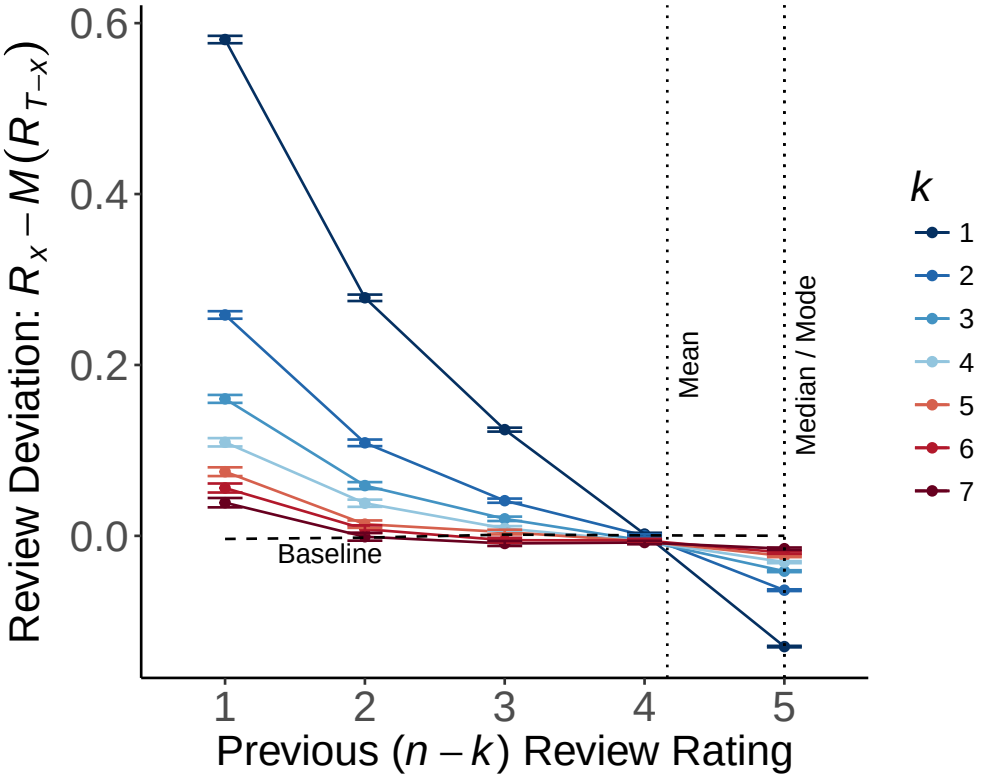


Figure 6.4: Deviation of the current review rating from the reviewer's average rating (y-axis) in relation to their previous review rating (x-axis) at k Review Distances for Amazon reviews

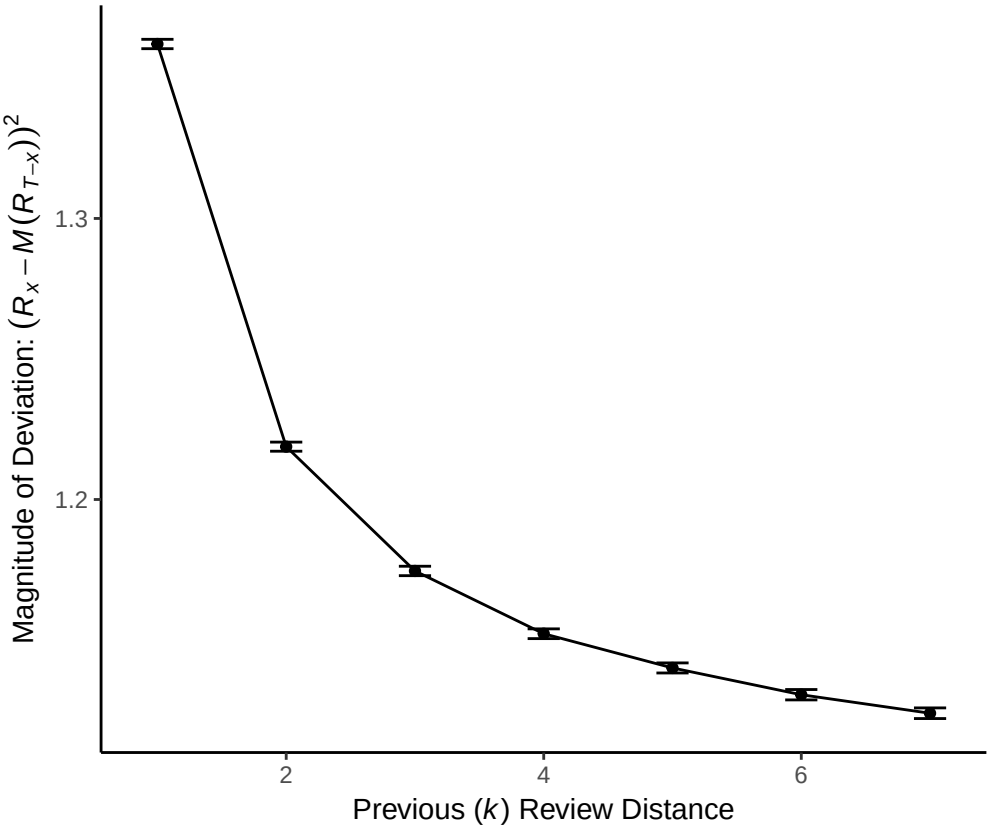


Figure 6.5: Magnitude of deviation of the current review rating from the reviewer's average relative to their previous review rating at k distances

effect on the current review, $F(1, 1.1x107) = 1.3x10^4$, $R^2 = .001$, $CI = (-.04, -.04)$, $p < .001$ such that as k gets larger, the magnitude of the effect of the previous review decreases Figure 6.5.

6.4 Discussion

We evaluated the presence of SDs in online review ratings across two different platforms and found subtle but significant dependencies. In fact, while the SD effects are indeed small in each set, they are astonishingly similar in both Yelp and Amazon. They support the predictions guided by laboratory experiments in categorization tasks (Zotov et al., 2011) and moral judgments (Olivola & Sagara, 2009). In both online review datasets, current ratings deviated from the average rating in contrast to preceding ratings: If a reviewer's previous rating was positive, their current rating is more likely to be less positive than their average rating. In addition, the contrast pattern was asymmetric. For example, a poor experience (1-star rating) with a restaurant made the subsequent restaurant appear more positive, and this upwards contrast was more powerful than the downward impression of a restaurant following an excellent preceding experience (5-star). However, this asymmetry may be the result of a positivity bias in reviews. The "even point" in our data appears to be 4-star reviews, where a prior review does not bias it. If this was centered on 3-stars, the observed contrast effect might appear more symmetric. Furthermore, the effect dissipates the farther the previous review is from the current review. There was no effect of previous review ratings when reviews were randomly ordered. Findings from

our study suggest that the observed contrast effects may be stable across reviewer contexts and products.

Note that the Yelp and Amazon review distributions are not normal, exhibiting a J-shape or bimodal distribution at 1-star and 4/5-stars with a mean of 3.75. Recent studies suggest that a J-shape bimodal distribution, unique to review data, may be the result of an underreporting bias (Hu et al., 2009), such that reviewers are more likely not to provide reviews when the average business rating is similar to their own experience. Interestingly, critics are more likely to have a unimodal distribution whereas non-critic reviewers tend to produce a J-shape distribution (Dellarocas, Narayan, et al., 2006).

Computational models that explain how sequential dependencies emerge from the decision-making process can help decontaminate current evaluations to produce a more accurate measure of one's experience (Mozer et al., 2010). Such models, though currently only developed for low-level perceptual tasks, might be fruitfully applied to areas such as online rating systems shown to impact a business's future success (Luca, 2016). Our current work is a step towards uncovering contamination effects that may be a rational property of the cognitive system within naturally occurring behavior. Developing psychologically-constrained tools that can adjust for such effects might help to provide ratings that better reflect a consumer's experience.

The Yelp and Amazon datasets indeed contain sources of noise. Reviewers sometimes don't review for various reasons and businesses change their names and their products adapt in real time to the demands of consumer behavior. SD experiments in the laboratory have control over the stimulus presented on each trial, but we lose this experimental control in the real world. Hence, there are several interpretations that we cannot rule out. For example, following a terrible experience with a restaurant, a rater may pay more attention to the selection process and as a consequence may actually go to a better restaurant the next time. The observed sequential dependency may be a change in selection behavior rather than a bias in decision. This is particularly why we believe that the approaches of laboratory experimentation and real-world data mining complement each other so well. The cognitive mechanism can be "captured" and studied in the controlled setting of the lab, and then "released" back into the real world, and we have reasonably good indications of what kinds of patterns to then search for in big data that reveal echoes of the cognitive system operating in the wild.

Research that tests laboratory-derived cognitive principles in the wild could also expand our understanding of these principles. The results here reveal that the effect size is smaller than might be expected, but the results also support the broader interest in sequential dependencies. These dependencies unequivocally appear in the noisy and nonstationary environment of human experience, in natural contexts such as evaluative activities during consumption (Olivola & Sagara, 2009). But the promise of this research goes beyond making the cognitive principles relevant to daily life. The data resources supplied by Amazon and Yelp will allow us to test the boundaries of these cognitive principles. Are there contexts in which SDs are weakened or even enhanced? For example, one-off experiences on Yelp such as highly expensive restaurants may be encoded in human memory quite differently from a restaurant experience that one might expect to have on a regular basis. These one-off experiences are unlikely to obliterate the SD, but they may be encoded in memory differently, and we might predict that they participant in somewhat weakened SDs. Such questions, outside the scope of the present paper, may be tested by examining connections among variables in the generous array of information afforded by these natural datasets. Future investigation of the structure of these datasets will allow researchers to search for these boundaries, and thus refine cognitive theory.

It is worth emphasizing that using cognitive principles with natural data can help practical endeavors in industry. Industry has become extremely focused on the importance of automated recommendation engines based on machine learning. These systems affect every facet of our daily lives, helping us to select options based on our previous preferences and global preferences across all individuals. But it is important to note that the upper limit on how well machine learning systems can perform is dictated by the quality of data provided by humans.

In our example, the human raters are not providing unbiased data from their experiences with the product or company. What is worse, they are probably not aware of the bias. Basing a recommendation system on sequentially contaminated data will be less than optimal, but for reasons that are not random. Hence, we reiterate the potential importance of the past century of experiments in psychological science to modern data mining enterprises.

Chapter 7

Discussion

The present work targets the growing need to understand what exactly is context and as a result elicits interesting questions about the nature of quantifying information. This comes at a time when more recent approaches in cognitive science are geared toward understanding the relationship between cognition and information in naturally occurring behaviors. This dissertation shows this can successfully studied through interdisciplinary collaboration; using tools from the field of computer science to mine increasingly larger datasets.

This chapter will review and summarize the projects presented here. I conclude by discussing their theoretical implications and future research.

The introduction (Chapter 1) begins by framing the current research as an investigation into the nature of information as it is shaped by context. It highlights how information, the statistical quantification of language, adapts to its context in an effort to more effectively communicate ideas. Further, the Information-Theoretic structure of language use is itself a context wherein micro-behaviors are constrained. Context not only constrains behavior, but helps to generate the types of behaviors that occur. These behaviors can be used to both invent and predict future events.

Chapter 2 explored how affective cognitive states relate to the underlying statistics of our language use. The relationship suggests that *how* we use language is influenced by the messages we intend to send (e.g., positive or negative). Being in an affective state —or simply writing an online review within the context of one —correlates positively with the information density of that review. In addition, there is a higher informational bandwidth in positive cognitive states. Theoretically, positive mood states consist of less cognitive noise. In practice, this means the use of information-dense terms, and the use of minimally information-dense terms occur in the same message. This can be explained when considering how function words are used to language.

Function words are very low information-dense terms. If we consider function words to be the floor, then less probable content words would not only increase the information density of reviews, but also increase message bandwidth. In the present study, negative reviews were high in information but low in Conditional Information Variability, whereas positive reviews were high in both. While negative and positive messages both express more information, positive reviews are able to do so using low information-dense terms. In situations of low cognitive noise, such as being in a positive mood state, function words may act as low information-dense frames that help prime the use of high information-dense terms; increasing both the amount of information and message bandwidth. This supports the hypothesis that when in a positive mood state, language users are likely to transmit information more easily than when in a negative mood state because they are able to use a wider range of information-dense terms to communicate their intended message.

This study provides a first step toward the use of larger datasets to explore the effects

of cognition in natural behavior as well as the use of machine learning models to predict review ratings using the statistical properties of the language used within them. Determining the success of statistical properties of language use in predicting behavior is one future direction.

Using a similar approach, Chapter 3 explored how the information density of reviews relates to the community in which they take place, by using theories from network science to quantify the structure of social networks in an online business review community. A simple cutoff method for building networks of a relative size was developed in order to determine how differences in network structures relate to differences in language use. The motivation for this work stems from findings that language acts more like a complex adaptive system, adjusting to various changes in its environment. Findings suggest that communities with more connectivity across its users are more likely to use information dense words. Users will often use more unique word combinations, suggestive that their communities afford more efficient idea sharing. Further, the variability of word use —a measure of the language user's channel to transmit ideas is also affected by simple network structures. This exploration suggests that local social interactions may contribute in interesting ways to language use. Further, the use of language is a highly complex, uniquely human skill that provides insight into online cognitive processes underlying its use. This study suggests that the space of possible cognitive states, and thus communicable ideas may dynamically adapt to both local and global constraints at multiple scales within a social community. A key strength of this work is the use of big-data. Without a large enough dataset, we would not have been able to determine differences in language use across many different types of communities. This approach affords uncovering new ways to test theoretical questions.

Chapter 4 takes a more practical approach. It stands as evidence of the process of science. My intentions were not to develop a new programming library, but to analyze large amounts of text data. After running into a computational bottleneck I was forced to tackle a more basic problem. This lead to the development of a new way to efficiently store and query large amounts of text data. There is a growing need for the development of scalable ways to process large datasets that cut across scientific disciplines. This process is advanced here by integrating efficiency tools from computer science into statistical programs already known by social scientists. By developing new packages in an easy to use statistical language such as `R` more efficient data processing tools are provided to more scientists. Not only does this new package broaden the number of tools social scientists can use to harness data, it also allows for very complicated `n`-gram computations to be calculated across millions of documents. This package is then put it to use where it is shown that even very complicated `n`-gram estimates of information relate to the language user's audience. More information dense language use was found to be more useful to its readers. From this speculate that more information dense messages communicate more ideas more efficiently.

In Chapter 5, language use is treated as a context wherein micro behaviors take place. Findings show the micro behaviors we use to produce language are related to the information density of their surrounding terms. That is, the very way we communicate, our dynamic micro-behaviors, are sensitive to the statistical properties of the language we use. Specifically, macro-level word-level frequencies influence the micro-level dynamics of the actions it takes to convey them. This suggests that language production is inherently proactive; influenced by contexual factors at a larger time scale. The context of one's current action is in part made up of anticipated future effects. It is the anticipation of the action's sensible effects which contribute to their dynamic unfolding in real time. Specifically, function words may be used as predictive frames for more information-dense content words. When the occurrence of a function word is less predictable, we produce that word at a slower rate. Our findings highlight the cognitive system's dynamic contextual dependence as a flow of information during mental processing. Even the most ballistic, rudimentary actions are constrained in real time by the statistics of our environment. This is contrast to accounts that describe language production as a ballistic sequence of events. Instead, motor actions were found to be sensitive to the distributional properties

communication at a broader scale.

Chapter 6, evaluated whether online review ratings were dependent on previous reviews. Theoretically, this work stands as a way to test how we might think about the influence that the previous behavior has on current actions. Previous chapters used previous behavior as the context to current behaviors. Similarly, this chapter assumes there is a statistical relationship between the occurrence of some previous behavior and current action in review ratings; however, the exact relationship is less well known. In fact, information theory is naive to the relationship that might exist. This chapter goes one step further by trying to uncover the exact relationship that exists between previous and current behavior. By stepping outside of language use this chapter studies the specific sequential dependencies of online review ratings, which unlike language, have a limited number of structured categories (e.g., 1-5). Findings indicate that current reviews ratings are sequentially dependent on previous review ratings across two large and independent databases. The effects of the previous review dissipates the farther away it is from the current review. The current findings validate assumptions and statistical models of information theory in language use as well as the effects of previous behavior on current behavior made in controlled laboratory settings. Using the most recent previous word as the current word's context captures at least some amount of the context's influence on the current word.

7.1 Future Directions

Both breadth and depth of this work could be extended in different ways to further study how context and information influence one another. In many cases, the observed effects are merely echoes of interesting cognitive effects on behavior and others appear broad and intuitive. I provide a few areas of ongoing research that aim to provide more depth and breadth to this research program while also highlighting the current work's limitations.

7.1.1 Optimal Behavior

One future direction may aim to test how word choice is influenced under efficiency constraints. When noise within a channel increases, the need to be more redundant in one's language use increases; lowering the average amount of information that can be transmitted reliably. Similarly, the amount of information within a message increases by reducing the amount of redundancy (Genzel & Charniak, 2002; Aylett & Turk, 2006; Jaeger & Levy, 2007; Jaeger, 2010). While the effects of fluctuations in the amount of noise present within a language user's channel have been well studied, the assumption, in all cases, is that the *length* of the message remains constant. However, it is often the case that conversations are cut short. Temporal constraints limit the length of messages that can be transmitted. Thus, information compression or loss must occur. What does language use look like when users are forced to convey their message with a limited number of words, or a limited amount of time? In such a case, two possibilities arise:

- The language user has already compressed the message as much as they can without incurring a penalty for additional compression, known as lossy information compression. The language user will sacrifice information density for reliability during message compression. This hypothesis is supported by theories that language users will increase redundancy in an otherwise information-dense message, essentially smoothing out their signal when it is too information-dense. Under this hypothesis, optimal behavior would be to sacrifice the amount of information transmitted while maintaining a relative rate of reliability.

- Alternatively, the language user might compress their message by removing redundant language, effectively increasing the average amount of information transmitted. In this

hypothesis, message compression is lossless as the language user will sacrifice the reliability (e.g., redundancy) of a message. As a result this increases the risk of being misunderstood. This would be done with the hope of transmitting the highest amount of information possible. Interestingly Chapter 4, found the language user's audience typically prefers information-dense messages, supporting this hypothesis of optimal behavior.

This study would test key theoretical assumptions about the way we communicate and the content contained within messages. If language users always choose to send reliable messages at the cost of sending less information, then content contained within messages would naturally be more dynamic; adjusting to the message's constraints. Alternatively, if language users sacrifice reliability for the chance to communicate a message in its entirety, then the content contained within a message is necessarily discrete and cannot be understood in segments. Testing these hypotheses would provide insight into optimal behavior under additional contextual constraints.

Experiment Decision making is influenced by the expected probability of the success of outcomes relative to their value. This is often considered in gambling games and quantified as the Expected Utility of a gamble:

$$EU = pA + (p-1)B \tag{7.1}$$

Where A and B are expected outcome values of a gamble with $n = \{A, B, .., \}$ number of outcomes and p is the probability of obtaining value A (for example).

Surprisingly, little work exists on how prior decisions influence the expected utility of the next decision. How does the probability of an expected outcome change given previous decisions? In the context of iterative binary decision making, EU can be considered a measure of the reliability of obtaining a given outcome. When provided with more than one binary decision, optimal decisions will use those choices in ways that maximize the reliability of their final decision. However, when only given a few choices, the most optimal use of those choices may be to try to maximize information gain at the cost of reliability. one possible expeiment is discussed below.

Participants are asked to locate an object on a grid of 64 cells (a standard chessboard) by asking only binary —yes/no —questions. For instance, participants might ask if the object is on a specific square, or if given more than one question, they may choose to ask questions that narrow down the object's possible location (e.g., "Is it on the right half of the board?"). In the context of information theory, each question has the potential of providing up to 6 bits of information. For example, if a participant asks if the object is on a single square, a "yes" answer would provide 6 bits of information: $-log_2(1/64) = 6$. A "no" answer provides significantly less since 63 of the 64 squares do not contain the hidden object: $-log_2(63/64) = .015$ bits. However, with more than one question the most optimal decisions will be those that aim to maximize the success of obtaining the correct answer. In this case, optimal questions increase the probability of getting the correct square on the last question. If given 6 questions, the optimal decision will be to ask questions that reduce the uncertainty of the object's location *by half* at each iteration: equal to 1 bit of information ($-log_2(1/2)$) per each question for a total of 6 bits, with an expected utility equal to 1 (100% reliability in discovering the location of the object).

When the number of choices decreases ($k - i$), the EU and reliability of one's final choice decreases. The most optimal decisions will be those that reduce the uncertainty in the final choice. That is, at no time should the individual choose to use all of their questions to guess individual squares. This would not maximize the probability of successfully locating the object and in doing so reduces the total possible information gain. Similarly, the decision maker would not choose to use their last question to continue reducing the number of possible choices if there are more than two (or more than 1 bit of information left to obtain). With the least number of

binary choices ($k = 1$), the optimal decision might be to ask a question that favors a possible outcome of 6 bits of information, by guessing the object's location to be on a single square.

Through the use of a decision making experiment, this study blends information theory and decision making while also providing a window into optimal behavior under temporal constraints. This stands to provide insight into the nature of language use under message constraints.

7.1.2 Topics

The structure of messages and channels in which they are communicated adapt to their current context. One future direction is to understand the content of messages, and how it might influence or constrain the way we communicate it (e.g., is it hard to talk about negative topics and easy to talk about positive ones?). Recent research has found incredible success in utilizing topic modeling, an unsupervised learning algorithm designed to uncover the *topics* across a set of documents. Further, understanding the topics within messages provide insight into the type of ideas the current context allows to be communicated. If so, we would expect to observe a relationship among topics and information.

7.1.3 Social networks

More natural social networks can be defined by selecting only those individuals who wrote reviews *within* the same business. This would (1) limit the number of reviews that each reviewer can provide (typically only one review per individual per business) and (2) provide a clear baseline between reviewers within networks and those who are not in networks.

Selecting businesses with many reviews (e.g., > 100) helps to ensure that a social network does exist within the business. It also provides a natural baseline (e.g., reviewers not connected to the business's social network). Further, this helps control the number of reviews provided by each individual and the topic of those reviews. Defining networks by business controls for differences in review ratings that individuals may have when reviewing different restaurants. If differences exist between network and no-network reviewers, specific aspects of ones network such as their overall connectivity, or cluster coefficient may shed light onto the structural aspects of one's network that are most likely to influence this behavior.

In addition to more controls, this study allows us to focus on how these networks come about. Given that each review is tagged with a date, it's possible to estimate when certain individuals became friends. This allows testing whether certain social network structures emerge under specific contextual constraints. Previous work shows that different social network structures are better adapted to different underlying statistical distributions (Goldstone, Roberts, & Gureckis, 2008). In this case, the topic, restaurant location or other features represent the network's statistical distribution and can be used to better understand its emergence.

7.2 Predicting Behaviors

Much of this work was motivated by theories and findings developed in closely controlled laboratory settings. As such, the practical applications of these theories are not well known. Many prediction tools have emerged in the last decade. There are now countless ways to test whether cognitive effects can predict natural behavior. Even subtle echoes provide insight into the underlying processes at work in natural and noisy behaviors. Unfortunately, such subtle effects and the way they were tested here are inherently retroactive. This dissertation is focused on finding the effects where they lie, but not in what can be predicted using these effects. A predictive application, in addition to merely pointing out that these effects exist, would

provide sound evidence of the merit of cognitive effects across many different disciplines including computer science, business, linguistics and more. It is possible to harness theories from cognitive and information science to predict behavior via feature engineering.

One approach may be to use a convolutional neural network, or deep neural network, to predict next best actions. *Next Best Action Marking* is a technique designed to provide businesses with suggestions on what actions they should take next to improve their business. Convolutional neural networks are designed to process data quickly by reducing the number of dimensions needed to accurately classify and predict missing data. By using this model in a novel way, cognitive, social and Information-Theoretic features of a business or review can be used to train the algorithm. After, it can be used to predict the value of additional variables of interest. For instance, a business may desire to have a higher review rating. To determine what needs to be changed to increase their star rating, the owner might input their desired star rating into the algorithm which will return a list of feature specific values related to restaurants with similar star ratings. This provides insight about the types of features that need to change for the business owner to increase their star rating (e.g., should the business have more reviewers, reviewers from the local social network, or perhaps different staff or food?).

The best machine learning algorithms were inspired by neural architecture. Computer scientists have been developing cognitively inspired machine learning models since the '80's, but have given little thought to engineering their features in similar ways. While a good algorithm should be able to take any type of natural data and process it effectively, using cognitively inspired features *and* algorithms may bring us closer to understanding how the mind processes the natural world.

7.3 Conclusion

The chapters of this dissertation detail the chronological process of science. The focus of the research program developed here is on understanding how ideas are shared, and how those ideas influence —and can be influenced by —the environment in which they take place. Our findings support a theory that even the most ballistic, reactionary behaviors are constrained and influenced by their contextual surroundings. The cognitive system utilizes the underlying statistical structure of its environment to make predictions about future states; perhaps in an effort to invent them via the successful transmission of ideas converted to actions that modify our physical and social surroundings.

The methods and tools we use to ask questions define the space of possible answers we can obtain. By providing tools, we expand the types of questions we can ask and their possible answers. This dissertation begins to describe a series of methods, tools, best practices and theories that the computational social scientist can use to harness the ever-changing ways we share ideas. Utilizing theories from cognitive science, tools from computer science and naturally occurring datasets, scientists can begin to navigate explore, uncover and predict cognitive, behavioral and information-theoretic effects in context. There may be "no universal cover of law", but progress in such a world depends on interdisciplinary methods and collaborations in an effort to maintain and, if lucky, accelerate the pace of scientific discovery.

References

Abney, D. H., Gann, T. M., Heutte, S., & Matlock, T. (Submitted). The language of uncertainty and political ideology of news sources in climate communication. *Cognitive Science*.

Abrams, D. M., & Strogatz, S. H. (2003). Linguistics: Modelling the dynamics of language death. *Nature*, *424*(6951), 900–900.

Angela, J. Y., & Cohen, J. D. (2009). Sequential effects: superstition or rational behavior? In *Advances in neural information processing systems* (pp. 1873–1880).

Aylett, M. (1999). Stochastic suprasegmentals: Relationships between redundancy, prosodic structure and syllabic duration. *Proceedings of ICPhS–99, San Francisco*.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31-56. doi: 10.1177/00238309040470010201

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, *119*(5), 3048-3058.

Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*(2), 179–190.

Baker, J. K. (1975). The dragon system–an overview. *Acoustics, speech and signal processing, IEEE transactions on*, *23*(1), 24–29.

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2015). Qrtengine: An easy solution for running online reaction time experiments using qualtrics. *Behavior research methods*, *47*(4), 918–929.

Baronchelli, A., Ferrer-i Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in cognitive sciences*, *17*(7), 348–360.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., ... Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language learning*, *59*(s1), 1–26.

Behmer, L., & Crump, M. J. (2015). Crunching big data with finger tips: How typists tune their performance toward the statistics of natural language. *M.. Jones (Ed.), Big Data in Cognitive Science*, 1–28.

Bentz, C., Verkerk, A., Kiela, D., Hill, F., & Buttery, P. (2015). Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLoS One*, *10*(6), e0128254.

Bentz, C., & Winter, B. (2013). Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change*, *3*(1), 1–27.

Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of experimental psychology: General*, *129*(2), 177.

Bonin, P., Chalard, M., Méot, A., & Fayol, M. (2002). The determinants of spoken and written picture naming latencies. *British Journal of Psychology*, *93*(1), 89–114.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & psychophysics*, *61*(2), 206–219.

Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language*, *34*(3), 383–398.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(3).

Cacioppo, J. T., & Gardner, W. L. (1999). Emotion. *Annual review of psychology*, *50*(1), 191–214.

Cantallops, A. S., & Salvi, F. (2014). New consumer behavior: A review of research on ewom and hotels. *International Journal of Hospitality Management*, *36*, 41–51.

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge University Press.

Cerni, T., Velay, J.-L., Alario, F.-X., Vaugoyeau, M., & Longcamp, M. (2016). Motor expertise for typing impacts lexical decision performance. *Trends in Neuroscience and Education*, *5*(3), 130–138.

Chater, N., Reali, F., & Christiansen, M. H. (2009). Restrictions on biological adaptation in language evolution. *Proceedings of the National Academy of Sciences*, *106*(4), 1015–1020.

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, *13*(4), 359–393.

Choi, H.-Y., Blumen, H. M., Congleton, A. R., & Rajaram, S. (2014). The role of group configuration in the social transmission of memory: Evidence from identical and reconfigured groups. *Journal of Cognitive Psychology*, *26*(1), 65–80.

Chomsky, N. (1975). *Syntactic structures*. Mouton de Gruyter.

Christakis, N. A., & Fowler, J. H. (2009). *Connected: The surprising power of our social networks and how they shape our lives*. Little, Brown.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and brain sciences*, *31*(5), 489–509.

Christiansen, M. H., Reali, F., & Chater, N. (2006). The baldwin effect works for functional, but not arbitrary, features of language. In *Proceedings of the sixth international conference on the evolution of language* (pp. 27–34).

Chua, K.-W., & Gauthier, I. (2015). Learned attention in an object-based frame of reference. *Journal of vision*, *15*(12), 899–899.

Chukharev-Hudilainen, E. (2014). Pauses in spontaneous written communication: A keystroke logging study. *Journal of Writing Research*, *6*(1), 61–84.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181–204.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. *Perspectives on socially shared cognition*, *13*(1991), 127–149.

Cohen Priva, U. (2010). Constructing typing-time corpora: A new way to answer old questions. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 43–48).

Cooper, W. E. (1982). *Cognitive aspects of skilled typewriting*. New York: Springer-Verlag.

Cormen, T. H. (2009). *Introduction to algorithms*. MIT press.

Cormode, G., & Muthukrishnan, M. (2011). Approximating data with the count-min sketch. *IEEE software*(1), 64–69.

Cormode, G., & Muthukrishnan, S. (2005). An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, *55*(1), 58–75.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, *14*(3), 391–403.

Dale, R., & Lupyan, G. (2012). Understanding the origins of morphological diversity: the linguistic niche hypothesis. *Advances in complex systems*, *15*(03n04), 1150017.

Davies, M. (2009). The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, *14*(2), 159–190.

Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, *95*(24), 14529–14534.

Dellarocas, C., Narayan, R., et al. (2006). A statistical measure of a population's propensity to engage in post-purchase online word-of-mouth. *Statistical Science*, *21*(2), 277–285.

del Prado Martın, F. M., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, *94*(1), 1–18.

Dey, A. K. (2001). Understanding and using context. *Personal and ubiquitous computing*, *5*(1), 4–7.

Diener, E., & Diener, C. (1996). Most people are happy. *Psychological science*, *7*(3), 181–185.

Dixon, P., McAnsh, S., & Read, L. (2012). Repetition effects in grasping. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *66*(1), 1.

Dodds, P. S., Watts, D. J., & Sabel, C. F. (2003). Information exchange and the robustness of organizational networks. *Proceedings of the National Academy of Sciences*, *100*(21), 12516–12521.

Donkin, C., Heathcote, B. R. A., & Brown, S. D. (2015). Why is accurately labelling simple magnitudes so hard? a past, present and future look at simple perceptual judgment. *The Oxford Handbook of Computational and Mathematical Psychology*, 121–141.

Doshi, A., Tran, C., Wilder, M. H., Mozer, M. C., & Trivedi, M. M. (2012). Sequential dependencies in driving. *Cognitive science*, *36*(5), 948–963.

Doyle, G., & Frank, M. C. (2015). Shared common ground influences information density in microblog texts. In *Hlt-naacl* (pp. 1587–1596).

Dryer, M. S., Gil, D., Comrie, B., Jung, H., Schmidt, C., et al. (2005). The world atlas of language structures.

Eddelbuettel, D., François, R., Allaire, J., Chambers, J., Bates, D., & Ushey, K. (2011). Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, *40*(8), 1–18.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Ember, C. R., & Ember, M. (2007). Climate, econiche, and sexuality: Influences on sonority in language. *American Anthropologist*, *109*(1), 180–185.

Estrada, C. A., Isen, A. M., & Young, M. J. (1997). Positive affect facilitates integration of information and decreases anchoring in reasoning among physicians. *Organizational behavior and human decision processes*, *72*(1), 117–135.

Everett, C. (2013). Evidence for direct geographic influences on linguistic sounds: the case of ejectives. *PloS one*, *8*(6), e65275.

Ferrer-i Cancho, R., Debowski, Ł., & del Prado Martín, F. M. (2013). Constant conditional entropy and related hypotheses. *Journal of Statistical Mechanics: Theory and Experiment*, *2013*(07), L07001.

Fine, A., & Jaeger, T. F. (2011). Language comprehension is sensitive to changes in the reliability of lexical cues. In *Proceedings of the cognitive science society* (Vol. 33).

Fine, A., Qian, T., Jaeger, T. F., & Jacobs, R. A. (2010). Is there syntactic adaptation in language comprehension? In *Proceedings of the 2010 workshop on cognitive modeling and computational linguistics* (pp. 18–26).

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, *47*(6), 381.

Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the cognitive science society* (Vol. 30).

Fredrickson, B. L. (2004). The broaden-and-build theory of positive emotions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*(1449), 1367.

Freeman, J., Dale, R., & Farmer, T. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*, 59.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, *1*(3), 215–239.

Freyd, J. J., & Finke, R. A. (1984). Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 126.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Furnham, A. (1986). The robustness of the recency effect: Studies using legal evidence. *The Journal of General Psychology*, *113*(4), 351–357.

Gagné, C. L., & Spalding, T. L. (2014). Typing time as an index of morphological and semantic effects during english compound processing. *Lingue e linguaggio*, *13*(2), 241–262.

Gandhi, S. P., Heeger, D. J., & Boynton, G. M. (1999). Spatial attention affects brain activity in human primary visual cortex. *Proceedings of the National Academy of Sciences*, *96*(6), 3314–3319.

Garner, W. (1953). An informational analysis of absolute judgments of loudness. *Journal of experimental psychology*, *46*(5), 373.

Gentner, D. R., Larochelle, S., & Grudin, J. (1988). Lexical, sublexical, and peripheral effects in skilled typewriting. *Cognitive Psychology*, *20*(4), 524–548.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 199–206).

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*. doi: 10.1177/0956797612463705

Goldstone, R. L., Roberts, M. E., & Gureckis, T. M. (2008). Emergent processes in group behavior. *Current Directions in Psychological Science*, *17*(1), 10–15.

Gordon, R. G., Grimes, B. F., et al. (2005). *Ethnologue: Languages of the world* (Vol. 15). sil International Dallas, TX.

Goyal, A., Daumé III, H., & Cormode, G. (2012). Sketch algorithms for estimating point queries in nlp. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 1093–1103).

Gray, R. D., & Atkinson, Q. D. (2003). *Language-tree divergence times support the anatolian theory of indo-european origin* (Vol. 426). Nature Publishing Group.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21 - 23. (The Changing Face of Cognition) doi: http://dx.doi.org/10.1016/j.cognition.2014.11.026

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *science*, *298*(5598), 1569–1579.

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable modified kneser-ney language model estimation. In *Acl (2)* (pp. 690–696).

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*(1), 11–26.

Holland, M. K., & Lockhead, G. (1968). Sequential effects in absolute judgments of loudness. *Attention, Perception, & Psychophysics*, *3*(6), 409–414.

Hsu, S.-M., & Yang, L.-X. (2013). Sequential effects in facial expression categorization. *Emotion*, *13*(3), 573.

Hu, N., Pavlou, P. A., & Zhang, J. (2006). Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of the 7th acm conference on electronic commerce* (pp. 324–330).

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the j-shaped distribution of product reviews. *Communications of the ACM*, *52*(10), 144–147.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain research*, *1626*, 118–135.

Isen, A. M., & Means, B. (1983). The influence of positive affect on decision-making strategy. *Social cognition*, *2*(1), 18–31.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23 - 62. doi: http://dx.doi.org/10.1016/j.cogpsych.2010.02.002

Jaeger, T. F. (2013). Production preferences cannot be understood without reference to communication. *Frontiers in psychology*, *4*.

Jaeger, T. F., & Levy, R. P. (2007). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems* (pp. 849–856).

James, W. (2013). *The principles of psychology*. Read Books Ltd.

Jastrow, J. (1900). The mind's eye.

Jelinek, F. (1976). Speech recognition by statistical methods. *Proceedings of the IEEE*, *64*, 532–556.

Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, *62*(S1), S63–S63.

Jesteadt, W., Luce, R. D., & Green, D. M. (1977). Sequential effects in judgments of loudness. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(1), 92.

Jones, M. N. (2016). *Big data in cognitive science*. Psychology Press.

Jurafsky, D., Chahuneau, V., Routledge, B. R., & Smith, N. A. (2014). Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, *19*(4).

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall.

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the linguistic inquiry and word count. *The American journal of psychology*, 263–286.

Kalish, Y., & Robins, G. (2006). Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure. *Social Networks*, *28*(1), 56–84.

Kandel, S., Peereman, R., Grosjacques, G., & Fayol, M. (2011). For a psycholinguistic model of handwriting production: Testing the syllable-bigram controversy. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(4), 1310.

Karl, F. (2012). A free energy principle for biological systems. *Entropy*, *14*(11), 2100–2121.

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, *22*(4), 751–761.

Kawamoto, A. H., Kello, C. T., Higareda, I., & Vu, J. V. (1999). Parallel processing and initial phoneme criterion in naming words: Evidence from frequency effects on onset and rime duration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 362.

Kawamoto, A. H., Kello, C. T., Jones, R., & Bame, K. (1998). Initial phoneme versus whole-word criterion to initiate pronunciation: Evidence based on response latency and initial

phoneme duration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(4), 862.

Kirsch, A., & Mitzenmacher, M. (2006). Less hashing, same performance: Building a better bloom filter. In *Algorithms–esa 2006* (pp. 456–467). Springer.

Klingenstein, S., Hitchcock, T., & DeDeo, S. (2014). The civilizing process in london's old bailey. *Proceedings of the National Academy of Sciences*, *111*(26), 9419–9424.

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Acoustics, speech, and signal processing, 1995. icassp-95., 1995 international conference on* (Vol. 1, pp. 181–184).

Knoblich, G., & Jordan, J. S. (2003). Action coordination in groups and individuals: learning anticipatory control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 1006.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790.

Krauss, R. M., & Fussell, S. R. (1990). Mutual knowledge and communicative effectiveness. *Intellectual teamwork: Social and technological foundations of cooperative work*, 111–146.

Kristjansson, A. (2006). Simultaneous priming along multiple feature dimensions in a visual search task. *Vision research*, *46*(16), 2554–2570.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606–608.

Labov, W. (1972a). *Language in the inner city: Studies in the black english vernacular* (Vol. 3). University of Pennsylvania Press.

Labov, W. (1972b). *Sociolinguistic patterns* (No. 4). University of Pennsylvania Press.

Laming, D. (1984). The relativity of 'absolute'judgements. *British Journal of Mathematical and Statistical Psychology*, *37*(2), 152–183.

Launchbury, J. (2017). A darpa perspective on artificial intelligence. *DARPAtv*.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... others (2009). Life in the network: the coming age of computational social science. *Science (New York, NY)*, *323*(5915), 721.

Levy, R. P., & Jaeger, T. F. (2006). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems* (pp. 849–856).

Li, M., Badger, J. H., Chen, X., Kwong, S., Kearney, P., & Zhang, H. (2001). An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, *17*(2), 149–154.

Liberman, A., Fischer, J., & Whitney, D. (2014). Serial dependence in the perception of faces. *Current Biology*, *24*(21), 2569–2574.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*(7163), 713.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the h&h theory. In *Speech production and speech modelling* (pp. 403–439). Springer.

Logan, G. D., & Crump, M. J. (2011). Hierarchical control of cognitive processes: The case for skilled typewriting. *Psychology of Learning and Motivation-Advances in Research and Theory*, *54*, 1.

Lohr, S. (2012). The age of big data. *New York Times*, *11*(2012).

Long, J. S. (1997). Regression models for categorical and limited dependent variables. *Advanced quantitative techniques in the social sciences*, *7*.

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using stata*. Stata press.

Luca, M. (2016). Reviews, reputation, and revenue: The case of yelp. com.

Lupyan, G. (2015). Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems. *Review of philosophy and psychology*, *6*(4), 547–569.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, *5*(1), e8559.

Lupyan, G., & Dale, R. (2015). The role of adaptation in understanding linguistic diversity. *Language structure and environment: Social, cultural, and natural factors*, 287–16.

Lupyan, G., & Spivey, M. J. (2010). Making the invisible visible: Verbal but not visual cues enhance visual detection. *PLoS One*, *5*(7), e11452.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

Markov, A. A. (1913). Primer statisticheskogo issledovanija nad tekstomevgenija onegina'illjustrirujuschij svjaz'ispytanij v tsep (an example of statistical study on the text ofeugene onegin'illustrating the linking of events to a chain). *Izvestija Imp. Akad. nauk*.

Martin, J. H., & Jurafsky, D. (2000). Speech and language processing. *International Edition*.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data. *The management revolution. Harvard Bus Rev*, *90*(10), 61–67.

McWhorter, J. (2002). What happened to english? *Diachronica*, *19*(2), 217–272.

Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in r. *Journal of statistical software*, *25*(5), 1–54.

Miall, R. C. (2003). Connecting mirror neurons and forward models. *Neuroreport*, *14*(17), 2135–2137.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, *26*(4), 393–424.

Mittlböck, M., Schemper, M., et al. (1996). Explained variation for logistic regression. *Statistics in medicine*, *15*(19), 1987–1997.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, *96*(2), 143 - 182. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0010027704001672` doi: http://dx.doi.org/10.1016/j.cognition.2004.09.001

Moreno, M. A., Stepp, N., & Turvey, M. (2011). Whole body lexical decision. *Neuroscience letters*, *490*(2), 126–129.

Mozer, M. C., Pashler, H., Wilder, M., Lindsey, R. V., Jones, M. C., & Jones, M. N. (2010). Decontaminating human judgments by removing sequential dependencies. In *Proceedings of the 23rd international conference on neural information processing systems* (pp. 1705–1713).

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful review? a study of customer reviews on amazon. com.

Mumma, G. H., & Wilson, S. B. (1995). Procedural debiasing of primacy/anchoring effects in clinical-like judgments. *Journal of clinical psychology*, *51*(6), 841–853.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142.

Nettle, D. (1998). Explaining global patterns of language diversity. *Journal of anthropological archaeology*, *17*(4), 354–374.

Nichols, J. (1992). *Linguistic diversity in space and time*. University of Chicago Press.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

Nowak, M. A., Komarova, N. L., Niyogi, P., et al. (2002). Computational and evolutionary aspects of language. *Nature*, *417*(6889), 611–617.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & psychophysics*, *60*(3), 355–376.

Nygaard, L. C., & Queen, J. S. (2008). Communicating emotion: linking affective prosody and word meaning. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(4), 1017.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46.

Olive, T. (2014). Toward a parallel and cascading model of the writing system: A review of research on writing processes coordination. *Journal of Writing Research*, *6*(2).

Olivola, C. Y., & Sagara, N. (2009). Distributions of observed death tolls govern sensitivity to human fatalities. *Proceedings of the National Academy of Sciences*, *106*(52), 22151–22156.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science a crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530.

Pate, J. K., & Goldwater, S. (2015). Talkers account for listener and channel characteristics to communicate efficiently. *Journal of Memory and Language*, *78*, 1–17.

Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-language perspective on speech information rate. *Language*, *87*(3), 539–558.

Pennebaker, J. W. (1997). *Opening up: The healing power of expressing emotions*. Guilford Press.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, *71*, 2001.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(04), 329–347.

Pinet, S., Hamamé, C. M., Longcamp, M., Vidal, F., Alario, F., et al. (2015). Response planning in word typing: Evidence for inhibition. *Psychophysiology*, *52*(4), 524–531.

Pinet, S., Zielinski, C., Mathôt, S., Dufau, S., Alario, F.-X., & Longcamp, M. (2016). Measuring sequences of keystrokes with jspsych: Reliability of response times and interkeystroke intervals. *Behavior research methods*, 1–14.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and brain sciences*, *13*(4), 707–727.

Plummer, P., Perea, M., & Rayner, K. (2014). The influence of contextual diversity on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 275.

Qian, T., & Aslin, R. N. (2014). Learning bundles of stimuli renders stimulus order as a cue, not a confound. *Proceedings of the National Academy of Sciences*, *111*(40), 14400–14405.

Qian, T., & Jaeger, T. F. (2012). Cue effectiveness in communicatively efficient discourse production. *Cognitive science*, *36*(7), 1312–1336.

Ramscar, M. (2013). Suffixing, prefixing, and the functional order of regularities in meaningful strings. *Psihologija*, *46*(4), 377–396.

Ramscar, M., Shaoul, C., Baayen, R. H., & Tbingen, E. K. U. (2015). Why many priming results don't (and won't) replicate: A quantitative analysis. *Manuscript, University of Tübingen*.

Reali, F., Chater, N., & Christiansen, M. H. (2014). The paradox of linguistic complexity and community size. In *Evolution of language: Proceedings of the 10th international conference (evolang10)* (pp. 270–277).

Ross, B. H., Wang, R. F., Kramer, A. F., Simons, D. J., & Crowell, J. A. (2007). Action information from classification learning. *Psychonomic bulletin & review*, *14*(3), 500–504.

Roux, S., McKeeff, T. J., Grosjacques, G., Afonso, O., & Kandel, S. (2013). The interaction between central and peripheral processes in handwriting production. *Cognition*, *127*(2), 235–241.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663–12668.

Rumelhart, D. E., & Norman, D. A. (1982). Simulating a skilled typist: A study of skilled cognitive-motor performance. *Cognitive Science*, *6*(1), 1–36.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, *35*(4), 606–621.

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. In *Collaboration technologies and systems (cts), 2013 international conference on* (pp. 42–47).

Sapir, E. (2004). *Language: An introduction to the study of speech*. Courier Corporation.

Scaltritti, M., Arfé, B., Torrance, M., & Peressotti, F. (2016). Typing pictures: Linguistic processing cascades into finger movements. *Cognition*, *156*, 16–29.

Schmidt, S. (2009). Shall we really do it again? the powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90.

Shaffer, L. (1973). Latency mechanisms in transcription. *Attention and performance IV*, 435–446.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, *27*(4), 623-656. doi: 10.1002/j.1538-7305.1948.tb00917.x

Shih, S. S.-y. (2014). *Towards optimal rhythm* (Unpublished doctoral dissertation). Stanford University.

Simmering, J. (2013, January). *How slow is r really?* [Blog]. `http://www.r-bloggers.com/how-slow-is-r-really/`.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059–1074.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Song, H., Dharmapurikar, S., Turner, J., & Lockwood, J. (2005). Fast hash table lookup using extended bloom filter: an aid to network processing. *ACM SIGCOMM Computer Communication Review*, *35*(4), 181–192.

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, *15*(5), 207–211.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(29), 10393–10398.

Stevens, S. S. (1957). On the psychophysical law. *Psychological review*, *64*(3), 153.

Stewart, N., Brown, G. D., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 3.

Stewart, N., Brown, G. D., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological review*, *112*(4), 881.

Stine, R. A. (1995). Graphical interpretation of variance inflation factors. *The American Statistician*, *49*(1), 53–56.

Stoll, S., Zakharko, T., Moran, S., Schikowski, R., & Bickel, B. (2015). Syntactic mixing across generations in an environment of community-wide bilingualism. *Frontiers in psychology*, *6*.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., & Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition*, *1*(2),

147–165.
Triandis, H. C. (1994). Culture and social behavior.
Trudgill, P. (1989). Contact and isolation in linguistic change. *Language change: Contributions to the study of its causes*, *43*, 227.
Trudgill, P. (2011). *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford University Press.
Valian, V., & Coulson, S. (1988). Anchor points in language learning: The role of marker frequency. *Journal of memory and language*, *27*(1), 71–86.
Van Galen, G. P. (1991). Handwriting: Issues for a psychomotor theory. *Human movement science*, *10*(2-3), 165–191.
Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and erp components. *International Journal of Psychophysiology*, *83*(2), 176–190.
Vilhena, D. A., Foster, J. G., Rosvall, M., West, J. D., Evans, J., & Bergstrom, C. T. (2014). Finding cultural holes: How structure and culture diverge in networks of scholarly communication. *Sociological Science*, *1*, 221–239.
Vinson, D. W., Abney, D. H., Amso, D., Chemero, A., Cutting, J. E., Dale, R., ... others (2016). Perception, as you make it. *Behavioral and Brain Sciences*, *39*.
Vinson, D. W., & Dale, R. (2014a). An exploration of semantic tendencies in word of mouth business reviews. In *Science and information conference (sai), 2014* (pp. 803–809).
Vinson, D. W., & Dale, R. (2014b). Valence weakly constrains the information density of messages. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual meeting of the cognitive science society* (p. 1682-1687). Austin, TX.
Vinson, D. W., & Dale, R. (2016). Social structure relates to linguistic information density. In M. Jones (Ed.), *Big data in cognitive science: From methods to insights.* Taylor & Francis.
Vinson, D. W., Dale, R., & Jones, M. N. (in revision). Decision contamination in the wild: Sequential dependencies in online review ratings. *Psychological Science*.
Vinson, D. W., Dale, R., Shih, S. S., & Spivey, M. (submitted). Statistics in the fingertips: Typing revealsword predictability. *Cognition*.
Vinson, D. W., Dale, R., Tabatabaeian, M., & Duran, N. D. (2015). Seeing and believing: Social influences on language processing. In *Attention and vision in language processing* (pp. 197–213). Springer.
Vinson, D. W., Davis, J. K., Sindi, S. S., & Dale, R. (2016). Efficient n-gram analysis in r with cmscu. *Behavior research methods*, *48*(3), 909–921.
Ward, L. M., & Lockhead, G. (1971). Response system processes in absolute judgment. *Attention, Perception, & Psychophysics*, *9*(1), 73–78.
Wilder, M., Jones, M., & Mozer, M. C. (2009). Sequential effects reflect parallel learning of multiple environmental regularities. In *Advances in neural information processing systems* (pp. 2053–2061).
Will, U., Nottbusch, G., & Weingarten, R. (2006). Linguistic units in word typing: Effects of word presentation modes and typing delay. *Written Language & Literacy*, *9*(1), 153–176.
Wolpert, D. M., & Flanagan, J. R. (2001). Motor prediction. *Current biology*, *11*(18), R729–R732.
Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, *117*(3), 543–578.
Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, *72*, 37–48.
Yoshimi, J., & Vinson, D. W. (2015). Extending gurwitsch's field theory of consciousness. *Consciousness and cognition*, *34*, 104–123.
Zipf, G. K. (1949). Human behavior and the principle of least effort: An introduction to human ecology.

Zotov, V., Jones, M. N., & Mewhort, D. (2011). Contrast and assimilation in categorization and exemplar production. *Attention, Perception, & Psychophysics*, *73*(2), 621–639.

Zwaan, R. A., & Pecher, D. (2012). Revisiting mental simulation in language comprehension: Six replication attempts. *PLoS ONE*, *7*, e51382. doi: 10.1371/journal.pone.0051382