

Copyright
by
Patrick Schultz
2017

**The Dissertation Committee for Patrick Schultz Certifies that this is the
approved version of the following dissertation:**

Gender Variation in Writing: Analyzing Online Dating Ads

Committee:

Lars Hinrichs, Supervisor

Mary E Blockley

Katrin Erk

Jacqueline M Henkel

Gender Variation in Writing: Analyzing Online Dating Ads

by

Patrick Schultz, BA, MA

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2017

Dedication

This dissertation is dedicated to my family: my parents Susanne and Christoph, my little sisters Fiona, Alison, and Catriona, my grandmothers Erika Fritz and Marianne Schultz, and my godmother Barbara Wittemann.

Acknowledgements

I would like to acknowledge my dissertation supervisor, Lars Hinrichs, without whose feedback and ideas this project would not have been possible. Thank you for your support throughout this process. I wish to thank my committee members, Mary Blockley, Jacqueline Henkel, and Katrin Erk, who were more than generous with their expertise and time. Thank you for serving on my committee. I would also like to thank everyone at the Digital Writing and Research Lab at the University of Texas at Austin, especially program coordinator Will Burdette, for kindly letting me use their lab facilities and equipment. Many thanks to all of you!

Gender Variation in Writing: Analyzing Online Dating Ads

Patrick Schultz, Ph.D.

The University of Texas at Austin, 2017

Supervisor: Lars Hinrichs

This dissertation presents a study of gendered language variation and linguistic indexicality in computer-mediated communication. A two-pronged approach combining the analysis of language production in a corpus of 103,000 English-language online dating ads with a language perception study (891 participants) is taken towards identifying the usage patterns and social meanings of nine features of e-grammar (Herring 2012). The indexicalities of features exhibiting gendered patterns in production as well as perception, emoticons (e.g. :) and prosodic items (e.g. *haha*), are discussed in light of their linguistic and social context. Drawing on empirical research on American gender ideologies, the study argues that they index characteristics such as friendliness and emotional expressiveness, both stereotypically associated with women. In an instance of indirect indexicality (Ochs 1992), they are then linked to femininity in this type of computer-mediated communication. In production, the same features exhibit a strong audience effect (Bell 1984): women, for instance, use them more frequently in ads directed at other women.

Throughout the analysis, the study makes use of and illustrates use cases for computational tools such as machine learning algorithms or automatic part-of-speech tagging in sociolinguistic research. At the same time, it attempts to strike a balance

between a quantitative, data-driven approach and the nuanced analysis of gender identities and linguistic indexicality in the performance of gendered identities.

Table of Contents

List of Tables.....	xi
List of Figures.....	xiv
Chapter 1: Introduction.....	1
1.1. Introduction.....	1
1.2. Study Outline.....	1
1.3. Background.....	2
1.3.1. Computational sociolinguistics.....	2
1.3.2 Language and gender.....	5
1.3.4. Contribution.....	7
Chapter 2: Language and Gender Research.....	8
2.1. Introduction.....	8
2.2. Early Studies.....	9
2.3. Gender in Quantitative Sociolinguistics.....	11
2.3.1. Labov’s “Gender Paradox”.....	12
2.3.2. Resolving the gender paradox.....	19
2.3.3. Gender in quantitative sociolinguistics: summary.....	25
2.3.4. Alternative approaches.....	26
2.3.5. Problems with the gender model in quantitative sociolinguistics.....	28
2.4. Gender and Language as Social Practice.....	31
2.4.1. Eckert: ‘The whole woman’.....	31
2.4.2. Eckert & McConnell-Ginet: ‘Think Practically, Look Locally’.....	34
2.4.3. Recent Research.....	38
2.5. Lakoffian Language and Gender Studies.....	44
2.5.1. The dominance model.....	45
2.5.2. The difference model.....	46
2.5.3. Linguistic representation of women.....	47
2.6. Gender Variation in Writing.....	48
2.7. Implications for this Study.....	52
Chapter 3: Production.....	54
3. 1. The Research Question.....	54

3.1.1. Previous Research.....	55
3. 2. The Dataset.....	59
3.3. Method.....	63
3.3.1. Feature extraction.....	64
3.3.2. Discovery.....	81
3.3.3. Analysis.....	93
3.4. Results.....	110
Chapter 4: Perception.....	112
4.1. The Research Question.....	113
4.1.1. Previous literature.....	114
4.2. Method.....	118
4.2.1 Creating the questionnaire.....	120
4.2.2. Creating the control stimulus.....	124
4.2.3. Creating treatment stimuli.....	126
4.2.4. Method evaluation.....	127
4.3. Results.....	128
4.3.1 Results control stimulus.....	129
4.3.2. Results treatment stimuli.....	130
4.3.3. Controlling for genre.....	130
4.3.2. Participant meta-commentary on gender-linked features.....	139
4.3.3. Dynamics of perceived author gender and perceived author attributes.....	142
4.3.4. Summary.....	147
Chapter 5: Conclusion.....	149
5.1. Summary: Results of Chapters 3 and 4.....	149
5.1.1. Chapter 3: production.....	149
5.1.2. Chapter 4: perception.....	152
5.2. Introduction.....	154
5.3. Integrating Production and Perception.....	154
5.3.1. Points of divergence.....	154
5.3.2. Points of convergence.....	155
5.4. Indexicality of Emoticons and Prosody.....	157
5.4.1. Genre conventions.....	158

5.4.2. Addressee effects	161
5.4.3. Gender ideologies	166
5.4.3. Social meaning of emoticons and prosody	176
5.5. Indexicality of Non-gendered Features	180
5.5.1. Capitalization and repeated punctuation	181
5.5.2. The gendering of Standard forms	185
5.6. Perceptual Attractiveness of E-grammar Features	189
5.7. Method Evaluation	190
5.7.1. E-grammar features in sociolinguistic theory	190
5.7.2. Perceptual salience	192
5.8. Conclusion and Implications.....	193
Appendix	196
A.1. Clustertools sample output	196
A.2. word2vec semantic groups.....	200
A.3. Perception survey questionnaire.....	251
Works Cited	255

List of Tables

Table 1:	Number of ads and words by gender in the dating ad corpus.....	60
Table 2:	Number of ads and words by category in the dating ad corpus.	61
Table 3:	Number of ads and words by addressee gender in the dating ad corpus..	61
Table 4:	Features of e-grammar, adapted from Herring (2012).	65
Table 5:	Ten most common acronyms and alphabetisms in the dating corpus.	69
Table 6:	Instances of non-Standard capitalization patterns in the dating corpus. ..	70
Table 7:	Ten most commonly capitalized tokens in the dating corpus.....	70
Table 8:	The most common items in camel and Pascal case in the dating corpus. .	71
Table 9:	Counts of clippings in the dating corpus.....	72
Table 10:	Counts of emoticons in the dating corpus.....	73
Table 11:	Counts of Leetspeak items in the dating corpus.....	74
Table 12:	Counts of items preceding <i>4</i> replacing <i>for</i> in the dating corpus.	75
Table 13:	Counts of items following <i>4</i> replacing <i>for</i> in the dating corpus.....	75
Table 14:	Counts of items preceding <i>2</i> replacing <i>to</i> in the dating corpus.	76
Table 15:	Counts of items following <i>2</i> replacing <i>to</i> in the dating corpus.....	76
Table 16:	Counts of items preceding <i>2</i> replacing <i>too</i> in the dating corpus.	77
Table 17:	Counts of items following <i>2</i> replacing <i>too</i> in the dating corpus.....	77
Table 18:	Counts of non-Standard punctuation in the dating corpus.....	78
Table 19:	Counts of prosodic items in the dating corpus.....	79
Table 20:	Counts of single letter replacements in the dating corpus.....	80
Table 21:	Word counts in the dating corpus.	81
Table 22:	Results of k-means clustering: cluster size and cluster membership by ad category.	90

Table 23:	Clusters as percentage of dataset and breakdown by category.....	93
Table 24:	Functions of micro-level features.....	101
Table 25:	Clusters as percentage of dataset and breakdown by category.....	102
Table 26:	Results of k-means clustering, distribution of categories over clusters in percent.	103
Table 27:	Semantic groups indicative of each cluster.	107
Table 27:	continued.	108
Table 28:	Questions assessing author perception in the matched guise study: type of question, response options, and perceptual variable tested for.	121
Table 28:	continued.	122
Table 29:	Linguistic tokens added to or changed in the control stimulus to create the six treatment stimuli.....	127
Table 30:	Questions assessing author perception in the matched guise study: response count for multiple choice questions, means for Likert questions. Results for genre-consistent dataset in parentheses.	133
Table 30:	continued.....	134
Table 31:	Perceived author gender in the matched guise study, percentages by stimulus. * indicates differences to control stimulus significant at the $p < 0.05$ level.	137
Table 32:	Perceived Craigslist category in the matched guise study, percentages by stimulus.	138
Table 33:	Masculine- and feminine-associated characteristics in the U.S. Adapted from Broverman et al (1972:70). Asterisks indicate features directly relevant to the present study.....	169
Table 34:	Most frequent responses to the “Women are ...” question using potentially primed attributes.....	172

Table 35:	Most frequent responses to the “Men are ...” question using potentially primed attributes.....	172
Table 36:	Most frequent responses to the “Women are ...” question using non-primed attributes.....	173
Table 37:	Most frequent responses to the “Men are ...” question using non-primed attributes.....	174
Table 38:	Most frequently capitalized words in ads from the m4m category.....	182
Table 39:	Most frequently capitalized words in ads from the w4w category.....	182
Table 40:	Most frequent non-Standard punctuation patterns in the m4m category.....	184
Table 41:	Most frequent non-Standard punctuation patterns in the w4w category.....	184

List of Figures

Figure 1:	Papers published on language and gender, 1996-2014 (Thomson Reuters 2015).....	8
Figure 2:	Negative concord in Philadelphia (Labov 2001:265).....	15
Figure 3:	PCA analysis of the Sydney speech community (Horvath 1985:71).....	27
Figure 4:	Direct and indirect indexes (Ochs 1992:342).....	37
Figure 5:	Frequency of features in the dating ad corpus, male versus female authors.	95
Figure 6:	Ranked feature frequency by category.....	97
Figure 7:	Relative feature frequency by category.	99
Figure 8:	Cluster membership by category, distance to value expected in random distribution in percent.....	105
Figure 9:	Perceived author gender by stimulus: responses to emoticons stimulus and prosody stimulus, compared to control stimulus.	136
Figure 10:	Full dataset: Means of perceived author characteristics, relative to control stimulus.	143
Figure 11:	Genre-consistent dataset: Perceived author characteristics, relative to control stimulus.	144
Figure 12:	Perceived author education by stimulus: responses to clipping stimulus and prosody stimulus, compared to control stimulus.	146
Figure 13:	Relative feature frequency by category.	151
Figure 14:	Genre-consistent dataset: Perceived author characteristics, relative to control stimulus.	153

Chapter 1: Introduction

1.1. INTRODUCTION

As of December 2016, 88 percent of the adult population in the United States use the internet (Pew Research Center 2017); 69 percent are members of an online social network such as Facebook, and 72 percent own a smartphone that allows them to use online messenger apps such as *WhatsApp* (Greenwood, Perrin & Duggan 2016). All of these figures have risen steadily over the last years. From a linguistic perspective, this increasing pervasiveness of computer-mediated communication (CMC) presents an intriguing new research opportunity. In sociolinguistics especially, CMC has attracted interest as a site of linguistic innovation and a novel social space for linguistic interaction (McKay 2011). Scholarly debates around CMC have touched on issues such as the mechanics of linguistic innovation in CMC (Danescu-Niculescu-Mizil et al. 2013), the social and linguistic consequences of a shift from spoken towards written interaction (Baron 2002), and novel ways of creating identities online (Danet 1998).

1.2. STUDY OUTLINE

This study enters the conversation about language variation in CMC by using a corpus of online personal ads to address an issue central to sociolinguistic research since its inception: linguistic gender differentiation. An overview of previous research on this issue is given in chapter 2. An analysis of 103,000 online personal ads in combination with a linguistic perception study, presented in chapter 3 and chapter 4 respectively, is used to investigate variation in use and the social meaning of nine items of e-grammar (Herring 2012), a set of linguistic features notable in CMC. In chapter 5, the study proceeds to investigate smaller-scale patterns in the use of these features, and explores their indexical

value. This includes a discussion and empirical analysis of American gender ideologies, genre conventions of dating ads, and of linguistic addressee effects evident in the data.

1.3. BACKGROUND

The study thus presents an approach to studying linguistic variation and linguistic indexicality in writing. To this end, it takes a two-pronged approach to analysis, combining a production and a perception study to take into account both sides of meaning-making in writing: the author *and* the reader of a text. The quantitative analysis is complemented by detailed, but data-driven account of the social and linguistic context. The goal of the analysis is to understand how linguistic variation serves to produce gender differentiation and to analyze the way specific linguistic indexicalities are activated in the context of the present dataset.

The quantitative analysis relies heavily on computational data mining and parsing in order to process text at scale. This computer-assisted approach situates the study within the field of computational sociolinguistics, an emerging, multi-disciplinary research agenda that attempts to marry two subfields of linguistics: sociolinguistics and computational linguistics. The goals and methodology of computational sociolinguistics are outlined below in section 1.3.1, followed by a description of the present study's methodological and theoretical underpinnings in language and gender research (section 1.3.2.), and a discussion of its potential contributions to the field in section 1.3.3.

1.3.1. Computational sociolinguistics

In their review of the computational sociolinguistics research agenda, Nguyen et al. (2016:4) define computational sociolinguistics as an

emerging research field that integrates aspects of sociolinguistics and computer science in studying the relation between language and society from a computational perspective.

(Nguyen et al. 2016:4)

Computational sociolinguistics, that is to say, attempts to apply the large-scale, data-driven methods developed by computational linguists to the study of language variation; at the same time, sociolinguistic insights are used to improve, add to, and challenge these tools (Nguyen et al. 2016:1). This integrated approach can thus draw on the empirical rigor and cutting-edge technology of computational linguistics and the detail-oriented and context-aware methodology developed by sociolinguists. Notable studies within the computational sociolinguistics framework include Bamman et al. (2014), Nguyen (2014), and Eisenstein (2017), all working on social aspects of language use on social media.

As a study of written language, the research presented here particularly benefits from the toolkit offered by computational linguistics, a research paradigm that works with text almost exclusively and has developed an extensive set of techniques to analyze written data at scale. Sociolinguists, on the other hand, have over the last 40 years developed sophisticated methodologies and a theoretical framework to systematically analyze linguistic variation, but have prioritized the study of speech over writing (Lillis 2013). Looking at the results of a computational analysis from this sociolinguistic angle will allow for examination of linguistic processes in more detail than is typically aimed for in computational “big data” studies, while still benefiting from the empirical power of a large dataset.

In this vein, the present study employs computational techniques for data collection, feature extraction, and analysis while keeping a sociolinguistic focus on

describing and understanding the linguistic behavior of speakers within their social worlds. Compared to a purely computational study, this sociolinguistic focus entails research design decisions such as working with a few, rather than thousands of linguistic features; making sure to extract all tokens of an item, rather than abstracting away from niche cases to build a predictive model that performs well across different settings; and attempting the step from the “how” to the “why”: if we establish that a group of people uses a certain linguistic variant, why do they do so in this context?

In practice, this means that computational tools, namely several Python and R scripts, were used to collect and clean the data: more than 103,000 dating ads from the website www.craigslist.org, containing a total of more than 10 million words. In order to compile as large a dataset as possible, all ads available at the point of data collection were downloaded and stored. To ensure data quality, concerns about genre effects (Herring & John C. Paolillo 2006), unreflected labeling for gender (Eckert 1990; 2014), and the researcher’s impact in general (Labov 1972) raised in previous sociolinguistic studies had to be addressed. As a consequence, the dataset focuses on one genre: online dating ads. Gender labels were not assigned by the researcher; rather, writers self-label for gender by posting in a specific category (such as “women for men”). Finally, the fact that the data was collected and generated outside of a laboratory or interview setting helps to minimize the researcher’s impact on the linguistic performance, what Labov (1972:209) called the “observer’s paradox”.

During the analysis, techniques common in computational linguistics such as regular expression pattern matching or part of speech-tagging were used to extract feature counts. During the analysis, machine-learning algorithms such as k-means clustering helped identify patterns in the data. Metrics developed in computational

linguistic research, such as the “term frequency – inverse document frequency” ratio, are used to evaluate each feature’s relevance in the dataset.

Insights from this computational text analysis, which is the subject of chapter 3, are complemented by the results of a matched guise language perception study presented in chapter 4. The matched guise technique (Lambert et al. 1960) is used in sociolinguistics to assess the perceptual value of linguistic forms. The perception study was designed to match the production study in linguistic as well as social context as closely as possible while maximizing the number of participants (891 participants total) to increase its potential of yielding statistically reliable results.

1.3.2 Language and gender

On the theoretical side, this study positions itself within – and tries to add to – a long history of language and gender research (outlined in chapter 2) from Lakoff (1973) to Bamman et al. (2014). The present study’s place within and potential contribution to this research agenda are sketched below.

Sociolinguists increasingly work with a social-constructivist understanding of gender, conceptualizing gender as a performance rather than a trait. (See section 2.4.). This performative model replaces the male-female gender binary with a dynamic understanding that allows for a variety of masculinities, femininities, as well as identities defying these categories. A crucial concept for this research is the notion of indexicality, or the social meaning of features (Silverstein 2003). Its premise is that linguistic features can index different stances, qualities, or other social constructs. The social meaning of a specific variant in a given context, however, is hard to pin down: Eckert (2008a) introduced the idea of an “indexical field”, comprising the potential social meanings of a linguistic form. The variants of a variable like word-final (ing), for instance, can index

stances such as “relaxed” versus “formal” as well as characteristics such as “educated” versus “uneducated”. Which of these meanings will be associated with the linguistic item in any context, Eckert (2008a:466) writes, “will depend on both the perspective of the hearer and the style in which it is embedded”.

In view of these theoretical developments regarding the definition of gender, the present study tries to allow for a more nuanced conception of the concept by using self-assigned gender labels. In addition, a clustering algorithm is used to identify relevant groupings by linguistic criteria alone (Bamman, Eisenstein & Schnoebelen 2014) instead of shoehorning the data into pre-defined gender categories. Incidentally, this moves the focus from a male – female model to an analysis that looks at gendered dyads (author gender plus addressee gender) as the relevant social groupings in this context. Finally, the study does not set out with a set of pre-conceived linguistics features hypothesized to be gender-linked. Instead, a set of nine features frequent in computer-mediated communication is used, and each member of the set analyzed for gendered meaning by inspecting its usage patterns and subjecting it to the perceptual evaluation task. That is to say, the general approach is to let relevant features emerge out of the dataset rather than imposing them on the data.

In order to understand the social concepts that surround and help create gender, the perception study was designed in a way to elicit as broad as possible an evaluation of features, rather than just asking for the perceived author gender.

Previous research on indexical meaning typically relies on researcher intuition in establishing which social meaning is activated in studies of production (e.g. Bucholtz 1999a) or on participant reaction in language evaluation studies (e.g. Campbell-Kibler 2008). The present study, as outlined above, combines the two approaches in a two-pronged study design, where the same features, occurring in the same context, are studied

in production and in perception. This approach helps illuminate both sides of the meaning-making process: the writer's performance and the reader's perception. In a second step, the study attempts to establish which social meaning is activated in the present dataset by considering its context, such as language ideologies, genre effects, and linguistic accommodation between author and addressee. Again, this analysis of social and linguistic context relies as much as possible on empirical findings from the two studies presented here as well as external research. It introduces data-driven accounts of language ideologies, genre effects and author-addressee interaction.

1.3.4. CONTRIBUTION

In the course of this study, a large dataset of online dating ads is compiled. This dataset is used to test a set of features, not yet explored in the language and gender literature, for potential gender-linked meanings. The paper's main methodological innovation is an attempt at combining the search for linguistic patterns in speaker production with an investigation into the perceptual relevance of the features, keeping factors such as feature set, locale, and genre constant. This amounts to integrating the classic variationist study (Labov 1966) with a matched guise study of language perception (Lambert et al. 1960) in order to paint a more comprehensive picture of the meaning of a linguistic form – the question, what does this feature mean to whom (Agha 2007)? Insights from these two studies inform the discussion of indexicality that follows and which similarly attempts to rely as little as possible on intuitions by empirically testing for abstract concepts such as gender stereotypes. At the same time, the study explores ways to productively use computational tools in a sociolinguistic study, such as clustering analysis to identify relevant patterns in the data.

Chapter 2: Language and Gender Research

2.1. INTRODUCTION

This chapter provides an overview of the research on language and gender relevant to the study presented in the chapters to follow.

The field of language and gender studies is relatively new and still expanding. The publication statistics in the Web of Science, a publication aggregator, for instance, show a steady increase in papers on language and gender within linguistics over the last twenty years. This figure includes all papers published in linguistics journals that included the words “language and gender” in title or abstract.

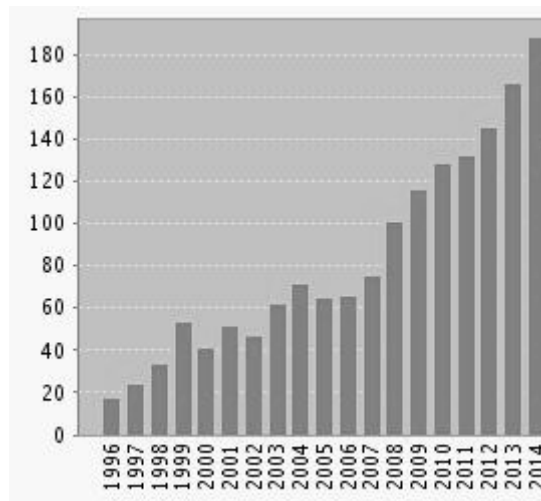


Figure 1: Papers published on language and gender, 1996-2014 (Thomson Reuters 2015)

With more than 180 papers published in 2014 alone, this outline by necessity focuses on the studies most directly relevant to the present study. For a more detailed overview, the reader is referred to recently published monographs outlining theory (Zimman, Davis & Raclaw 2014; Litosseliti 2006; Mills 2012), and methodology (Baker 2014; Mills &

Mullany 2011; Harrington et al. 2008) of language and gender research, as well as a wide array of introductory textbooks on the topic (Ehrlich, Meyerhoff & Holmes 2014; Eckert & McConnell-Ginet 2013; Pichler & Coates 2011; Litosseliti 2006; Sunderland 2006; Coates 2004). Studies of language and gender also continuously appear in journals such as *Language Variation and Change*, *Language in Society*, the *Journal of Sociolinguistics* or *Gender and Language*.

2.2. EARLY STUDIES

The earliest research addressing gender-linked language differences was done by anthropologists compiling lists of gender-exclusive word forms among indigenous tribes. Rochefort (1666) lists different words used by men and women in his *Caribbean (sic) Vocabulary*. Henry (1879) describes gender-specific affixes in the grammar of the Chiquitos in Bolivia; Dixon & Kroeber (1903) find gender-exclusive words among the Yana in California. More detailed linguistic descriptions include Haas (1944) on the Koasati in Louisiana, Sapir (1949) on gender-specific forms used by the Yana, and Bradley (1988) on the Yanyuwa in Australia. Coates (2004:28–34) gives a detailed overview of the hallmark studies of this anthropological research.

The first linguistic study of gender differences in English is Jespersen's (1964) chapter on "The woman", first published in 1922. Covering putative gender differences in phonetic features, word choice and speech processing, Jespersen (1964:245) finds very few gender differences in the pronunciation of words. Analyzing gender variation in word choice, Jespersen (1964:245) postulates a tendency of women to use euphemisms and other polite forms while men prefer direct and "often rude" denominations. At the same time, men are mainly responsible for creating new words (Jespersen 1964:247). In general, according to Jespersen (1964:248) women have a smaller vocabulary than men but are

better at learning and using language. Jespersen's supports his claims with evidence mostly from novels, personal observation and linguistic commentary in the media. Jespersen (1964:251) concludes that all the gendered features he identified are "only preferences that may be broken in a great many instances and yet are characteristic of the sexes as such". Jespersen argues, that is, that English has gender-*preferential* rather than gender-exclusive forms. This approach informed all the studies of gender differences on English pursued afterwards.

While Jespersen was among the first to explicitly address this kind of gender variation, tacit assumptions about female and male language use informed the studies of his contemporary dialectologists as well. They often operated under the presumption that women used more innovative word forms than men. Dialect geographers such as Gilliéron (1902) in France, Bartoli (1927) in Italy and Orton (1962) in England sought out as informants what later has been called NORMs: non-mobile, older rural males (Chambers & Trudgill 1998:29). Orton (1962:15), editor of the *Linguistic Atlas of England*, for instance, writes: "In this country men speak vernacular more frequently, more consistently, and more genuinely than women". Informant statistics presented in Coates (2004:37) show that female speakers are accordingly under-represented in the early dialect atlases. Of Orton's 989 informants, twelve percent were women; nine percent of Gilliéron's French and 13 percent of Bartoli's Italian informants were female.

This shows both that linguists from the beginning believed that there was some kind of gender difference in language use and that their findings reflected and shaped popular conceptions about gender differences.

This early research on gender differences often tried to establish whether the observed differences were due to biological differences, that is: nature, or result of the social context, that is: nurture. In linguistics, gender-related research from very early on

perceived any differences as socially conditioned (Cameron 2009). Evidence for biological differences in language processing remained slim (Macaulay 1978a; Philips 1987:6) and the sociolinguistic framework in which most of the research was done unsurprisingly yielded mainly social explanations. Very closely related to the nature – nurture debate is the terminological differentiation between sex and gender. Traditionally, these terms were defined as sex relating to the biological binary, gender to the social binary built on it (Eckert & McConnell-Ginet 2013:2). Early sociolinguistic studies used the term interchangeably (Eckert 1990) and more recent work often treats what is called “biological sex” as socially constructed as gender is (Eckert & McConnell-Ginet 2013:4). It will be seen that the nature – nurture debate turns up very rarely in sociolinguistics (see e.g. section 2.3.2.3) and has failed to generate much interest. The controversy is considered settled in favor of social over biological causation. This paper thus will focus on gender as a socio-culturally constructed binary (Cameron 2009).

2.3. GENDER IN QUANTITATIVE SOCIOLINGUISTICS

Systematic investigation into correlations between gender and language use started with the rise of quantitative sociolinguistics, initiated by the early studies of Labov (1966; 1972). Gender differentiation became one of the most thoroughly studied aspects of linguistic variation. The results concerning gender are “among the clearest and most consistent” in quantitative sociolinguistics, Labov (1990:205) writes. In his review of the literature, Chambers (2009:116) summarizes these findings as follows:

In virtually all sociolinguistic studies that include a sample of males and females, there is evidence for this conclusion about their linguistic behavior: women use fewer stigmatized and non-standard variants than men do of the same social group in the same circumstances.

This pattern was described in the classic quantitative studies, such as Trudgill's (1974) study of Norwich English, Wolfram's (1969) research in Detroit and Labov's (1966) New York City study.

Accordingly, the theorem is widely accepted in the field: statements to the same effect as Chambers' above are found in the major introductory textbooks such as Wardhaugh & Fuller (2014:172), Coulmas (2013:48), Meyerhoff (2011:207), Tagliamonte (2012:32), Shilling (2011:223), Trousdale (2010:67), Trudgill (2000:70), or Hudson (1996:193) as well as in books aimed at the general public (Edwards 2013:107).

2.3.1. Labov's "Gender Paradox"

The most sophisticated account of the principle described above – women use less non-standard or stigmatized forms – is given in Labov's monograph on the *Principles of Linguistic Change* (2001), building on ideas first formulated in Labov (1990). In Labov's list of principles of linguistic change, the "linguistic conformity of women" (Labov 2001:266) is Principle 2. In addition to this Principle 2, Labov includes two more principles relevant to gender: Principle 3, which states that "[i]n linguistic change from above, women adopt prestige forms at a higher rate than men" (Labov 2001:274)¹, and Principle 4, which asserts that "[i]n linguistic change from below, women use higher frequencies of the innovative forms than men do" (Labov 2001:292). As noted above, Principles 2, 3, and 4 are parts of a longer list of general rules of linguistic change that Labov puts forward in his book. The other principles (such as Principle 1, the "curvilinear Principle") are not relevant to the topic matter of this paper, but for ease of reference,

¹ In the book, this statement about women and change from above is indeed introduced as Principle 3 (Labov 2001:274). It should not be confused with a different and unrelated Principle 3: the "Principle of uniform evaluation" (Labov 2001:214).

Labov's original numbering is used here. The three principles relevant to gendered language use are described in more detail below.

2.3.1.1. Labov's Principle 1: Women use more prestige variants

Regarding Principle 2, the linguistic conformity of women, Labov writes: "For stable sociolinguistic variables, women show a lower rate of stigmatized variants and a higher rate of prestige variants than men" (Labov 2001:266). Labov argues that the finding about women using more high prestige variants than men is consistent and reliable: "Principle 2 is a strong and broad generalization" (Labov 2001:271).

This dynamic is observed in variables including for instance the use of [ɪn] instead of [ɪŋ] in words like *walking*: regarding this variable, studies by Labov in Philadelphia (Labov 2001:264–65) and New York City (Labov 2006), Trudgill (1974) in Norwich and Fischer (1958) in a New England all find female speakers to use the prestige variant of /ɪŋ/ more often than men. Another English sociolinguistic variable deemed stable by Labov that has been studied in some detail is the realization of the interdental fricatives /θ/ and /ð/ as a stop or affricate such as [d̪] or [t̪]. Again, women are found to use the stigmatized forms less often than men when it comes to /θ/ and /ð/ in Philadelphia (Labov 2001:264–65), New York City (Labov 2006), Belfast (Milroy & Milroy 1978), Detroit (Shuy 1968) and among African-American speakers in Detroit (Wolfram 1969).

Similar findings are reported for negative concord (as in *I don't know nothing*), where the low-prestige variant of multiple negation is found less often in the speech of women in Philadelphia (Labov 2001:264–65), New York City (Labov 2006), Anniston, Alabama (Feagin 1979), among African-American speakers in Detroit (Wolfram 1969) and in Detroit (Shuy 1968).

Other variables show a similar stratification. The realization of post-vocalic /r/, for example, a high-prestige variant in Mainstream U.S. English, is found to be used more often among female speakers in Detroit by Wolfram (1969). The deletion of copula (*He busy right now*), a feature of AAVE, is less common among female speakers in Detroit (Wolfram 1969). Macaulay (1978b) studied nonstandard vowel realizations among Glasgow school children and found girls to use more standard, prestige forms than boys. Milroy & Milroy (1978) found realization of several vowels and intervocalic /th/ to be strongly gender-stratified in various Belfast neighborhoods. Nichols (1983) found younger women to consistently use fewer creole forms in a Gullah-speaking community in South Carolina. Meshtrie et al. (2015) find that female speakers of South African English tend to use a more Standard variant of the BATH vowel.

Results that do not conform to this principle thus far have been reported from some Muslim societies, e.g. in Nablus, Israel, where Jawad (1987) found men to use the /qaf/ prestige form of Classical Arabic more frequently than women. Bakir (1986) and Sallam (1980) also found male speakers to use standard forms more frequently in Iraqi Arabic and Egyptian Arabic respectively. This reversal of the usual findings has been attributed to local gender norms or to a mis-interpretation of the concept of prestige (Jawad, cited in Labov (2001:270)). Ibrahim (1986), for instance, argues that standard, classical Arabic is not a prestige variety of Arabic but a self-contained variety of its own.

However, in all studies that do find the gender pattern described above – women using more prestige forms than men – this behavior is strongly stratified by social class: Labov's (2001:265) data on negative concord in Philadelphia illustrates this.

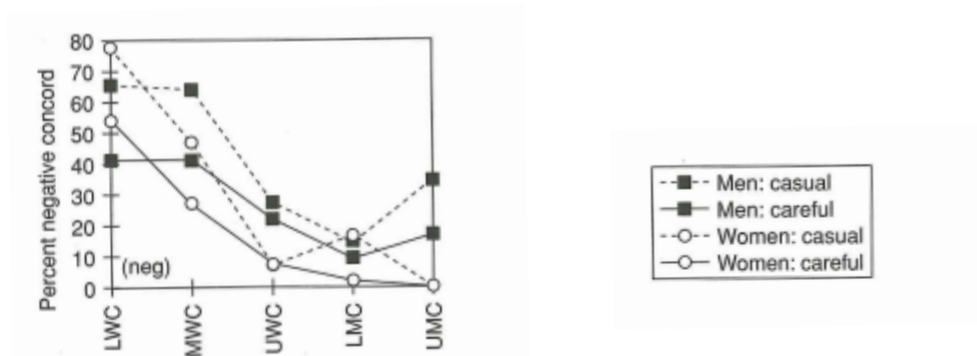


Figure 2: Negative concord in Philadelphia (Labov 2001:265)

Looking at the middle-working class (MWC), for example, we see that that female middle working class speakers do indeed use the low-prestige variant less often than middle working class men do: 49 percent versus 65 percent in casual speech.² But if we compare the same females to speakers of other social groups, we find that speakers of the upper-working class, the lower-middle class, etc. use the low prestige form even less, irrespective of their gender. Thus, social class seems more relevant than gender in explaining this variable: it is only within their own socio-economic class that females consistently score higher than males on the prestige-scale for this feature.

Put more briefly: Principle 2 holds true for most speakers *once we control for social class*.

Another confounding factor in these results is style (in Labov’s terminology, this refers to “careful” speech, as in a reading passage, versus “casual” speech, as in a sociolinguistic interview). For the stable variables described above, the rate of the high-prestige variants is consistently higher in “careful” speech (Eckert & Rickford 2001). This is not surprising if we, following Labov, assume that speakers in this context pay more attention to their performance and try to speak “correctly”. Looking again at the effect of

² Labov’s middle working class includes speakers ranked from 4–6 on the socioeconomic class index described in Labov (2001:58–73).

style in Labov's (2001:265) data on negative concord above, we see that middle-working-class women use significantly less negative concord in careful speech. However, this lower number still exceeds the usage rates for upper-working class or middle class speakers in *casual* style. Thus, even when these middle-working class speakers pay close attention to their speech, they still use more low-prestige forms than the "casual" middle-class speakers.

Labov (2001:265) describes this as "an intimate and complex interaction between style, gender and social class". This complex interaction will be relevant later on in this chapter.

Labov's other two gender-relevant principles, Principle 3 and Principle 4, contrast with Principle 2 in that they apply to sound changes in progress rather than established variables. Labov distinguishes between two types of linguistic change: changes from below (which operate "within the system, below the level of social awareness", Labov (2001:279)) and changes from above (which "take place at a relatively high level of social consciousness", Labov (2001:274)). His Principle 3 establishes gender differences in such changes from above, Principle 4 discusses changes from below.

2.3.1.2. Labov's Principle 3: Women lead change from above

Principle 3 describes the role of gender in change from above: "In linguistic change from above, women adopt prestige forms at a higher rate than men" (Labov 2001:274). Labov (2001:274) writes that women lead the acquisition of new prestige patterns as well as the elimination of stigmatized forms. Examples of such changes from above include the shift towards the pronunciation of postvocalic /r/ in New York City (Labov 2006), vowel changes in French-speaking Montreal (Kemp & Yaeger-Dror 1991), changes in a Native American language of Labrador (Clarke in Denning (1987)), vowel raising in Belfast

(Milroy & Milroy 1978) and the loss of rural dialect features in Spain (Holmquist 1985). A similar logic applies to language shifts, such as one from Hungarian to German in an Austrian town analyzed in Gal (1978), where women are leading the move away from the heritage language. A similar development is described in Nichols' (1983) study of Gullah speakers in the U.S. Regarding the interaction of gender with other social factors in change from above, Labov (2001:275) writes: "The interaction of sex and social class that was found for stable sociolinguistic variables is even more characteristic of changes from above".

2.3.1.3. Labov's Principle 4: Women lead change from below

Principle 4, then, describes the role of gender in the second type of linguistic change, change from below: in these more sub-conscious changes, women use higher frequencies of innovative forms than men do, Labov (2001:292) argues. Female speakers lead those changes, just like they do in changes from above. Evidence for Principle 4 comes from studies as early as Gauchat (1905), who found women to lead sound changes in Charmey, Switzerland. Later, Labov (2006) describes women in New York City leading the way in several sound changes. Trudgill (1974) identified the same mechanism in sound changes in Norwich English. In a later study of Philadelphia English, Labov (2001, p. 289) finds women leading 13 out of 16 vowel changes in progress. Clarke et al (1995:214) find women leading a vowel shift in Canada. Analyzing data from the *Atlas of North American English* (Labov, Ash & Boberg 2006), Labov (2001:288) argues that women lead this vowel shift: "For the NCS as a whole, women are in advance of men". Eckert (1989) describes a similar dynamic in her study, as does Britain (1992) in his study of high-rise-terminals in New Zealand. Tagliamonte & D'Arcy (2004) document a pragmatic change (the use of *be*

like) that is led by female speakers. Changes in progress thus, according to Labov (2001:283), show

a consistent majority pattern of women leading men. [...] In none of these cases do we see the creation of a stable sex differentiation. Rather, the mechanism of the change crucially involves the initiating role of women at the outset, and the later adoption of the change by men.

However, a few studies also find men to be leading a sound change, e.g. Labov's (1972) analysis of sound change on Martha's Vineyard, the un-rounding of (o) in Norwich (Trudgill 1974) or the rounding of /a/ in Belfast (Milroy & Milroy 1978).

2.3.1.4. The gender paradox

Building on three principles described above – that women tend to use less non-Standard forms, that they adopt prestige variants quicker in change from above and lead in change from below – Labov finds that the results of quantitative sociolinguistic research leave us with what he calls the “Gender Paradox”: “Women conform more closely than men to sociolinguistic norms that are overtly prescribed, but conform less than men when they are not” (Labov 2001:293).

In other words: in stable variation, women are conservative and conforming to prestige norms; in changes from above, they are progressive, that is moving towards the prestige norm. In leading change from below, however, they are non-conforming since these innovations are by definition neither overtly prescribed nor prestigious. We can thus rephrase the gender paradox as a conformity paradox: “Women deviate less than men from linguistic norms when the deviations are overtly proscribed, but more than men when the deviations are not proscribed” (Labov 2001:367). This phrasing will be relevant for some of the attempts at explaining the gender paradox presented below.

2.3.2. Resolving the gender paradox

While this gendered pattern is by now widely recognized in the literature, researchers are usually hard-pressed to explain it. An explanation based on female status-consciousness was presented by Trudgill (1972); Labov (1990) emphasized the role of women as caregivers; Chambers (2009) attempted a partly biological account; Gordon & Heath (1998) invoked sound symbolism, and Deuchar (1988) discussed the results in a politeness framework. Their arguments are outlined below.

2.3.2.1. Overt and covert prestige

Trudgill (1972:182), analyzing data from Norwich, argues that women tend to use more standard variants because they are more status-conscious than men. He proposes two reasons for this hypothesized difference. First, women are less secure in their social position and need to signal their status by linguistic means (Trudgill 1972:182). Second, Trudgill argues, women gain social recognition for how they *appear*, including their linguistic appearance, rather than what they *do*, which is the relevant metric for men (Trudgill 1972:183). These two reasons lead to women having to rely on linguistic status symbols, such as high-prestige forms, more than men. Trudgill also attempts to explain the higher frequency of non-standard features among male speakers, the other side of the coin by introducing the concept of “covert prestige” (Trudgill 1972:183). Building on Labov’s (2006) findings in New York City, Trudgill argues that working-class, non-standard forms have covert prestige (or “hidden value”, (1972:183)) for male speakers, since they carry connotations of masculinity, roughness, and toughness (Trudgill 1972:183). Note that Trudgill’s focus on prestige as the explanatory factor is, often implicitly, shared by most of the studies cited above. In their attempts at explaining variation, these studies often invoke the high prestige (or: desirability) of Standard forms.

2.3.2.2. Women as caregivers

Labov (1990) offers an explanation for the gender paradox based on traditional gender roles. He attributes the fact that women usually lead linguistic change to their role as primary caregivers (Labov 1990:243; 2001:307). Their advanced forms are picked up by children learning the language. Small children, Labov argues, just are not very often exposed to advanced male-dominated changes. Thus, the female-dominated innovations are passed on to the next generation while male-lead changes are not.

Regarding the “paradox behavior” of women being at the same time rule-conforming with stable variables and non-conforming with some ongoing sound changes, Labov (2001:374) in his Philadelphia data finds a positive correlation between advanced vowel forms and the use of established non-prestige forms among higher female socioeconomic groups (7-14 on his scale, i.e. the upper working and the middle class). This suggests that women who deviate from the norm in changes in progress tend to do so in stable variables as well – there is no paradox for these speakers. Rather, the presumed paradox reflects a “split within the female population” (Labov 2001:376): one female group is conforming all the time while the other female group tends to be non-conforming. Labov (2001:376) concludes: “Two different sets of women are involved. Or to put it another way, the leaders of linguistic change differ consistently from the rest of the population”. The gender paradox is thus, at least for Philadelphia, resolved as a mis-interpretation of statistical findings. Labov (2001:376) summarizes: the assumption about “women treating new sound changes differently from old ones and stable variables”, was an “error”.

2.3.2.3. Biological differences

Chambers (2009) offers a two-pronged explanation for gender differences in language. To some extent, he says, the stratification described in Labov’s linguistic gender principles

can be explained by gender-based differences: the different roles ascribed to men and women in the respective communities. *Pace* Labov, he also introduces the idea of sex-based variation, differences that can be due to biological differences (Chambers 2009:141).

When it comes to gender-based differences, Chambers focuses on social and geographical mobility as a function of gender roles, drawing mainly on research by Milroy & Milroy (1978) in Belfast and Wolfram (1969) in Detroit. “[T]he dynamic variable”, Chambers (2009:139) argues, “is mobility”: the breadth of social and geographical contacts influences a speaker’s use of linguistic variants. In the various settings investigated, Chambers argues, gender roles happen to make women more mobile than men. Thus, their linguistic interactions tend to be more diverse. All of this, according to Chambers, allows women to become acquainted with and acquire prestige forms from outside their own social group. This social mobility, Chambers (Chambers 2009:136) claims, can partially account for the women’s tendency to use more prestige – in these cases: non-local – features.

However, Chambers argues that these gender differences alone cannot account for the consistent gender-patterning presented above: if linguistic differences were just due to socialization, we would expect them to be more pronounced in speech communities with such different gender conceptions as African-Americans in Detroit, Irish working class speakers or Arabic speakers from Cairo. Instead, Chambers (2009:148) argues that the differences in linguistic behavior also reflect “sex-based variability”, that is to say biological differences. Women, he argues, command a wider range of linguistic forms and use a larger repertoire of styles because of a female “neuropsychological verbal advantage” (Chambers 2009:151, italics removed). Citing neurobiological research (Halpern 1986; Denno 1982; Maccoby & Jacklin 1974), Chambers states that women have an advantage

in verbal abilities over men and consistently perform better in studies of skills such as language fluency and comprehension, size of vocabulary and spelling skills (Chambers 2009:146).

The female tendency to use more high-prestige features is thus just one result of their verbal superiority, according to Chambers. It illustrates that women are better at adapting to new linguistic situations and more able to employ a wide range of linguistic features in general. Chambers is careful to point out that this linguistic advantage is a relative, rather than an absolute finding: not every woman is better than every man at using language, but the overall trend is strong enough to result in the statistical effects seen in the studies cited above.

Chamber's idea of sex-based variability was criticized by Romaine (1996:868) for its "mechanistic way" of approaching variation and for neglecting to sufficiently take cultural context into account. Labov (2001:276–77) presents some counter-evidence from neurobiological research and argues that any existing biological differences are too small to account for the major differences in sociolinguistic behavior. Labov (2001:277) also points out that if women were more capable of picking up on and manipulating linguistic resources, one would expect them to do better than men on linguistic self-evaluations (when asked what linguistic variants they typically use, for example), which they don't. James (1996:118) writes that the evidence for a biological female advantage in verbal ability is "very tenuous", citing a meta-study (Hyde & Linn 1988) that finds gender differences in verbal ability to be negligible. Philips et al (1987) concur.

2.3.2.4. Sound symbolism

Gordon & Heath's (1998) attempt at accounting for gender differences is based on theories of sound symbolism. Discussing the tendency of women to lead sound change,

Gordon & Heath posit that women only lead a specific kind of sound change: vowel fronting and raising, a feature of most of the vowel changes described above.

[W]e hypothesize that women are attracted to particular vocalic qualities, prototypically the high front unrounded vowel [ī], while men are attracted to other vocalic qualities, prototypically back vowels rounded or not, namely, [a ɔ o u]”.

(Gordon & Heath 1998:423).

This attraction, Gordon & Heath argue, is partly due to the general higher frequencies of female vowels – a function of their smaller vocal apparatus – and partly to sound symbolism. The idea of sound symbolism rests on the notion of the frequency code, an idea developed by Ohala (1994). Put simply, it states that low frequencies, e.g. in voice, are associated with size and strength; high frequencies, on the other hand, indicate smallness and weakness. Gordon and Heath see the frequency code at play in the universal tendency of women to prefer higher pitch and high vowels like [ī]. They hypothesize that this accounts for women leading shifts that involve fronting and raising, which could be interpreted as a long-term movement towards [ī]. Gordon and Heath’s theory explicitly includes only long solid-state vowels; other sounds such as diphthongs or consonants are disregarded. Their model has been criticized (Holmes & Britain 1998) for not accounting for all the data found in the literature and lumping together change from above and change from below. Labov (2001:291) presents counter examples and points out that this theory fails to explain consonantal changes which pattern similarly to vowels.

2.3.2.5. Politeness and power

Deuchar (1988) argues that language-external factors such as social class or status consciousness cannot satisfyingly account for gender variation. Instead, she proposes a

language-internal, pragmatic approach based on the politeness model developed by Brown & Levinson (1978). Two key notions of this model are 'face', the public self-image claimed by the speaker, on the one hand, and 'power', the power differential between speaker and addressee, on the other.

Applying Brown & Levinson's terminology to variationist studies, Deuchar (1988:31) argues that women, due to their lower social position, typically are powerless speakers. According to Brown & Levinson's politeness model, this means they have to pay a lot of attention to preserving the face of their addressee, for example by appearing non-imposing and by showing their approval of the interlocutor. At the same time, they have to protect their own face, their desire for the addressee to approve of them in return. Using standard linguistic variants is one way of achieving this, Deuchar (1988:31) argues:

[S]tandard speech, with its connotations of prestige, appears suitable for protecting the face of a relatively powerless speaker [i.e. a female speaker] without attacking that of the addressee.

Men, on the other hand, don't have to worry about their or others' face as much, since they are the powerful participant in a gender-mixed interaction (Deuchar 1988:30). In her review of Deuchar's model, James (1996:112) accepts that it "might be valid for at least some communities" but argues that it is mainly useful for explaining women's interactions with a Standard-speaking researcher (which would probably apply to most of the studies cited above). Deuchar's model does not account, James argues, for studies reporting that women still use more standard forms when talking to friends or a non-standard speaker (a result reported in Larson (1982) and Cheshire (1982)). Most importantly, some studies show women to use the same amount or less standard forms than men even when talking to a male standard speaker (Khan 1991; Rickford 1991; Thomas 1988). This is contrary to what Deuchar's model predicts.

2.3.3. Gender in quantitative sociolinguistics: summary

To summarize: early quantitative sociolinguistic studies share certain characteristics that influence their findings on gender.

First, these studies tried to identify mechanisms of linguistic change: their focus was on the role externalities (such as gender) played in linguistic change. The interest was in identifying whether women or men were “leaders of change”. The research paradigm was not all that interested in gender variation as such.

In addition to that, authors were mainly interested in describing patterns of variation along social class lines. They categorized speakers according to socio-economic class. This model of class stratification was their primary independent variable; results on gender-differentiation were merely a by-product (Lesley Milroy 1992:164).

When trying to explain variation along gender lines – or any other social axis – early sociolinguists relied heavily on “prestige” as the explanatory variable. Prestige was usually tied to social class: variants used among members of a higher socio-economic class were automatically assigned high prestige.

The type of change investigated most frequently was sound change. The literature overwhelmingly deals with vowel changes. Consonants, morphological or semantic features are under-represented. Variation in writing is under-studied for the same reason.

Ideologically, variationist sociolinguistics understands itself as an empirical research paradigm, analyzing language in a “dispassionate and accountable manner” (Lesley Milroy 1992:163). The research generally steers clear of political or social statements or activism.

2.3.4. Alternative approaches

Two studies (Horvath 1985; Milroy 1987) tried different, though still strongly quantitative approaches to linguistic variation. Both have important implications for the study of language and gender and are summarized below.

2.3.4.1. Gender and social networks

Milroy (1987) shows how linguistic gender differences can result from the different kinds of social networks that women and men typically engage in. In a study of sociolinguistic variation in working-class neighborhoods of Belfast (Milroy & Milroy 1978), Milroy (1987:123) finds that overall, men use more non-standard, vernacular variants than women. Milroy (1987:156) also establishes that men on average have denser, more close-knit social networks. These dense social networks, Milroy (1987:137) argues, function as norm-enforcement mechanisms: they help to maintain local vernacular speech forms among their members. For a lot of the variables she discusses, network strength emerges as a better predictor of variation than gender. This follows from her observation that when women do form dense social networks as the men above, they also use more vernacular variants: in one neighborhood, for instance, a group of young women were found to have denser networks than their male counterparts. Accordingly, they used more vernacular features than the men did (Milroy 1987:149).

The tendency of male speakers to use more vernacular variants (what Labov would call “low prestige” variants), Milroy’s study suggest, might be a result of their involvement in dense, local social networks. “[A] generalization based on the sex of the speaker rather than, for examples, his social values, or the structure of his social network, is unwise”, Milroy (1987:113) summarizes. Cheshire (1982) reports similar findings in her study of adolescents in Reading: Similar to the speakers from Belfast, male youth in

Reading engage in more tight-knit networks than the girls (Cheshire 1982:89). Boys also adhere more closely to vernacular norms and values.

2.3.4.2. Clustering linguistic variation

The second study to be discussed is Horvath's (1985) study of English in Sydney, Australia. She analyzed the realization of five vowels, four consonants, progressive (ing) and high rising terminal intonation.

Her approach could be described as the polar opposite of the standard Labovian study, which groups speakers by social criteria and then correlates these groups with use patterns of linguistic features. Horvath, on the other hand, used a statistical algorithm to cluster her data according to linguistic criteria (such as whether a certain vowel was backed or fronted) without taking any non-linguistic speaker characteristics into account. She then tried to correlate the resulting groups with the social characteristics of the speakers they included. In the graph reproduced below, for instance, speakers fall into two clusters according to their realization of the five vowels analyzed.

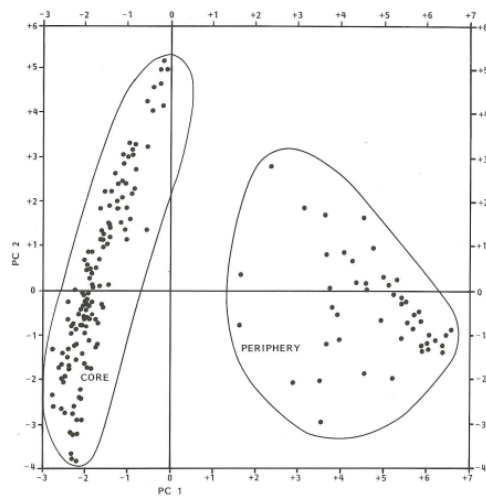


Figure 3: PCA analysis of the Sydney speech community (Horvath 1985:71)

“One of the most important features of this kind of analysis”, Horvath (1985:178) writes, “is that the quantitative description of a speech community is based solely on linguistic behavior, surely an appropriate starting place for a linguist”. Her approach, which does not initially categorize speakers according to social class, indicates that at least in Sydney gender is often a stronger predictor of variation than social class. “[T]he definition between males and females was the most marked, but socioeconomic class, age and ethnicity were all found to be important in understanding the linguistic variation in the speech community”, Horvath sums up her results (1985:174). She also finds that oftentimes consonants rather than vowels serve as gender makers (Horvath 1985:169).

She notes, however, that gender might be important because of the specific social situation in Australia, where class stratification is not as strong (it is often even called a “classless society”: Horvath (1985:4)) as in, for example, Britain.

2.3.5. Problems with the gender model in quantitative sociolinguistics

Most of the tenets of quantitative sociolinguistics listed above – the model of social stratification, the primacy of social class, the strongly empirical approach – proved controversial to some extent: summarized below are the critiques by Acker (1973) and Delphy (1981), who question the validity of the stratificational model as such; by Milroy (1989; 1992) who argues that “prestige” might not be a useful concept in explaining variation; and by Cameron & Coates (1985), who perceive Labov’s “dispassionate” approach to be inherently sexist. An overview of a more fundamental critique regarding gender specifically, independently formulated by Eckert (1990), Eckert & McConnell-Ginet (1992), Romaine (1996), James (1996), and Cameron (1998a), follows. Most of their criticism is succinctly summarized in Eckert (1990) and Eckert & McConnell-Ginet

(1992). These two papers are therefore discussed in detail below; some additional points brought up by other authors are introduced first.

2.3.5.1. Stratification

The stratification model used in the early studies by Trudgill and Labov was borrowed from sociology, where it was quickly criticized for its “intellectual sexism” (Acker 1973). This model of social stratification makes certain assumptions regarding gender and the social structure, outlined in Acker (1973:937), including that the family is the unit in the stratification system and that the social position of the family is determined by the status of the male head of the household. The status of women is thus determined by that of the males to whom they are attached. Some problematic implications of this approach to gender are given in Delphy (1981). She is especially critical of the practice to equate a woman’s social class – typically operationalized as occupation –with that of her husband. While women are categorized according to their own occupation while single, they are assumed to be members of their husband’s class once they are married, no matter their own employment status (Delphy 1981:114). This approach, Delphy argues, is inconsistent because it applies different criteria not only to men as opposed to women, but also married versus single women (Delphy 1981:119–20). The lack of methodological rigor might obscure findings and makes results difficult to compare across studies or populations.

2.3.5.2. Primacy of social class

Summarizing twenty years of quantitative research in the Labovian tradition, Milroy (1992:165) comes to a “depressing conclusion”: these studies failed to “greatly advance our understanding of the nature and role of sex differentiation in language, or how it interacts with social class variation”. She attributes this failure to the researcher’s relying

on social class as the main explanatory factor and treating gender as a variable of secondary importance. Referring to studies like Labov's analysis of negative concord in Philadelphia shown above, Milroy (1992:168) writes that "[.S]ex and class differentiation should not be locked into an inseparable nexus, with sex differences being explicated in terms of class differences". Milroy (1992:171) instead argues for the idea of a "sociolinguistic division of labour" where some variables serve to index social class, others gender. (Consonants, Milroy (1992:168) suggests, might be more likely to mark gender than class). James (1996) concurs and argues that the predictive power of gender has been under-utilized in sociolinguistics. "[T]here is now considerable evidence that socioeconomic class is not necessarily a more basic variable than sex or gender", James (1996:106-7) writes. James cites the studies by Horvath (1985) and Milroy & Milroy's (1993) social network studies to support this idea. Cameron & Coates (1985:146) also point out that "the possibility of norms that are sex *and* class specific is never entertained" in the work of early sociolinguists.

2.3.5.3. Prestige

Milroy (1989; 1992) argues that the concept of prestige is under-defined and often mis-applied in sociolinguistics. According to Milroy, prestige is too often directly tied to socio-economic status, assuming that higher social class-variants inherently have "prestige" for everyone. In a similar vein, Romaine (1982) and Milroy (1982) question the assumption that the Standard language is always the most prestigious form. Milroy (1989; 1992) also points out that sociolinguists often don't distinguish between institutional prestige (e.g. measured in income) and local prestige (e.g. measured in number of contacts). Especially when discussing gender-variation, Milroy (1989:221) argues the "prestige-based explanation [...] thus seems to have some serious flaws in it".

Other researchers such as Lakoff (1973) or Eckert (1990) argue that power, not prestige, is the relevant explanatory variable in linguistic gender differentiation. These arguments are discussed in more detail in section 2.4. below.

2.4. GENDER AND LANGUAGE AS SOCIAL PRACTICE

In addition to these methodological issues, some linguists argued that the entire sociolinguistic enterprise was in need of an overhaul when it came to doing gender-related research. Arguments to this effect were made by Eckert (1990), Eckert & McConnell-Ginet (1992), Romaine (1996), James (1996) and Cameron (1998a). Their main line of argumentation is succinctly summarized in Eckert (1990) and in Eckert & McConnell-Ginet (1992). These two papers are therefore discussed in detail in sections 2.4.1. and 2.4.2. below.

2.4.1. Eckert: ‘The whole woman’

Eckert (1990) argues that early sociolinguists like Labov oversimplified the concept of gender. For instance, they categorized speakers according to sex, thus substituting the biological concept of sex for the social construct of gender. Eckert writes that

differences in patterns of variation between men and women are a function of gender and only indirectly a function of sex [...] we have been examining the interaction between gender and variation by correlating variables with sex rather than gender differences.

(Eckert 1990:247)

This simplistic approach to gender differs remarkably from the sophisticated theories of social class often employed in the same studies, Eckert (1990:246) notes. Eckert points

out that feminist theory, unlike sociolinguistic theory, now sees gender as a construct rather than given. As elaborated on below, that means that “female” and “male” will mean very different things in different communities. This, Eckert argues, needs to be taken into account in sociolinguistic studies: for instance, linguists cannot expect gender categories to have the same linguistic effect across populations.

Thus, according to Eckert, previous sociolinguistic attempts to identify general gender differences in language that apply to entire populations, if not humankind, were bound to fail. Eckert (1990:247) writes that “there is no apparent reason to believe that there is a simple, constant relation between gender and variation”. Rather, other factors such as class, ethnicity, or context will strongly interact with gender. Briefly put, just because two speakers are considered part of the same gender group, they need not necessarily be very similar in their linguistic behavior. A female working-class speaker is not somehow inherently linguistically similar to a female upper-class speaker just because the researcher considers them both to be “female”. Rather, researchers need to understand gender as a form of social practice (Eckert 1990:253); that is, explore the construction of gender in the community studied, requiring some ethnographic fieldwork.

Echoing some of the thoughts brought up by Milroy in section 2.3.5.3 above, Eckert thinks that linguists need to move away from relating linguistic variants to prestige. Eckert argues that power, rather than prestige, is the important underlying sociological concept when discussing gender. The gender hierarchy, according to Eckert, is a power hierarchy (where men have and women don’t have power) and we cannot understand linguistic processes without referring to power at some point: “Above all, gender relations are about power and access to property and services” (Eckert 1990:256).

Eckert’s main theoretical point is that this power structure forces women into the accumulation of symbolic capital: in the traditional gender system, Eckert argues, men go

out to work and acquire financial and social capital. Women have to stay home and are “thrown into the accumulation of symbolic capital” (Eckert 1990:256). Eckert here invokes the notion of the linguistic marketplace, developed by Bourdieu (1975), where linguistic forms have a certain value attached to them. Accumulating this kind of symbolic capital – by dressing, acting, and speaking “right” – is often the only way a woman can gain status within her community.

This leads Eckert to hypothesize that women should show a broader range of linguistic variation and in the use of other indicators of group membership. This idea is borne out by data from a U.S. high school that she presents in the paper: she found girls to align themselves more strongly than boys with the linguistic conventions of their social group. This is evident, for example, in the fronting (or backing, depending on group association) of certain vowels. “[C]ategory membership”, Eckert (1990:256) summarizes “is more salient to members of one sex than the other; girls are asserting their category identities through language more than are the boys”. A noteworthy aspect of this finding is that Eckert’s High School girls cannot be said to speak more or less standard than the males – they actually occupy both ends of the spectrum, depending on the social group they are aligning with.

Similarly, Eckert’s study does not categorize her speakers according to occupation or the income level of their family like previous research would have done. Rather, Eckert tries to establish groups that are relevant to the speakers themselves: in this setting, the “Jocks” and “Burnouts”. This leads Eckert to stress that in order to understand gender variation, the researcher needs a comprehensive understanding of the social categories that are relevant to the speakers. Linguists need to establish the relevant social groups within the community (“Jocks” and “Burnouts” in her example), explore the notions and ideologies of gender (as pointed out above), and figure out the power structure that

governs social interactions (empowered boys and powerless girls in her study). Eckert does so through extensive ethnographic observation.

Eckert adds a few more theoretical points of interest. First, she suggests that linguists need to start perceiving gender as a continuum rather than a binary (Eckert 1990:247), just like some scholars already do for social class or age. She also notes that gender is different from other social categories in a very striking way: despite – or because of – all their differences, linguistic and otherwise, men and women are expected to “team up” with a member of the other group. This is unusual, Eckert writes:

It is not a cultural norm for each working-class individual to be paired up for life with a member of the middle class or for every black person to be paired up with a white person. However, our traditional gender ideology dictates just this relationship between men and women.

(Eckert 1990:253–4)

This makes gender roles what she calls “reciprocal”: differences are established and stressed to create a distance but are also intended to attract members of the other group.

2.4.2. Eckert & McConnell-Ginet: ‘Think Practically, Look Locally’

In a second paper, Eckert & McConnell-Ginet (1992) outline what they consider an appropriate research methodology for language and gender research. They argue that most previous research on language and gender fails to take context into account appropriately:

Citations abound in support of claims that women’s language reflects conservatism, prestige consciousness, upward mobility, insecurity, deference, nurturance, emotivity, connectedness, sensitivity to others, and solidarity; and that men’s language reflects toughness, lack of affect, competitiveness, and

independence. But the observations on which such claims are based have all been made at different times and different circumstances with different populations.

(Eckert & McConnell-Ginet 1992:485).

Their main piece of advice to address this issue is given in the paper's headline, which asks to "think practically and look locally". That is to say, the focus of research must be on linguistic practice; the object of interest is the local interaction in a specific community of practice (see below for a definition). Eckert & McConnell-Ginet (1992:472) argue that the crucial question needs to be "how gender is constructed in social practice, and how this construction intertwines with that of other components of identity and difference, and of language". This community-based practice orientation entails that first, gender cannot be separated from other aspects of other social identities and relations; second, that gender does not have the same meaning across communities; and third, that the linguistic manifestation of these meanings differ across communities, too. Eckert & McConnell-Ginet argue that one cannot separate linguistic performance from a speaker's general style: one also needs to pay attention to a speaker's way of dressing, behaving, interacting, etc. since all of this will also impact the perception of their linguistic performance. "Language", Eckert & McConnell-Ginet (1992:332) write, "is never the whole story".

Three concepts developed in Eckert & McConnell-Ginet's paper in particular have shaped sociolinguistic research in the years to follow: the community of practice and the notion of gender as performance, which in turn requires the concept of indexicality. Each is introduced in a short paragraph below.

The "community of practice" is Eckert and McConnell-Ginet's alternative to the "speech community" typically employed in earlier studies. Eckert and McConnell-Ginet (1992:463) write

A community of practice is an aggregate of people who come together around mutual engagement in an endeavor. Ways of doing things, ways of talking, beliefs, value, power relations – in short, practices – emerge in the course of this mutual endeavor.

This concept was developed by Wenger (1999). Its advantages for sociolinguistic studies are described in more detail by Bucholtz (1999b): among them the fact that the community of practice allows researchers to take practices other than language into account; that it allows them to focus on marginal community members as well as central ones; that the “Community of Practice” is less dependent on the researcher’s personal judgment and that it gives more agency to speakers.

In addition to the community of practice, the idea of gender as a performance is central to Eckert & McConnell-Ginet’s ideas. Rather than a static category, *gender* “becomes a dynamic verb” (Eckert & McConnell-Ginet 1992:462). Or, as it is often put: Gender is something you *do*, not something you *are*. This concept was developed in feminist theory, most famously by Butler (1999 [1990]), who writes that

Gender is the repeated stylization of the body, a set of repeated acts within a highly rigid regulatory frame that congeal over time to produce the appearance of substance of a natural sort of being.

(Butler 1999:43–4)

Butler’s framework sees language as one of the means to perform gender. This notion goes back to the work of Austin (1962) and his concept of the performative utterance; his idea that utterances are actions of some form is the cornerstone of performativity theory in linguistics (Hall 1999:184).

Finally, the concept of linguistic indexicality has been crucial in efforts to pursue this kind of research on gender performances. Indexicality is the relationship between a

linguistic form – be it lexical, phonological, syntactic, etc. – and its social meaning. Gender can be one of these social meanings. Extrapolating from the research above, for instance, we might argue that the pronunciation of (ing) as [ɪŋ] indexes female-ness. Indexing gender, however, is not as straightforward, as for example Ochs (1992) points out. Echoing the thought formulated by Eckert above, Ochs (1992:336–7) writes that the “relation between language and gender is not a simple straightforward mapping of linguistic form to social meaning of gender”. There are some direct linguistic indexes of gender, such as gendered pronouns, Ochs (1992:340) points out – but those are rare. Most of the time, we are dealing with indirect indexical relations: As the schema reproduced below indicates, linguistic forms can evoke (that is, index directly) certain stances or activities. These stances and activities are in turn linked to gender.

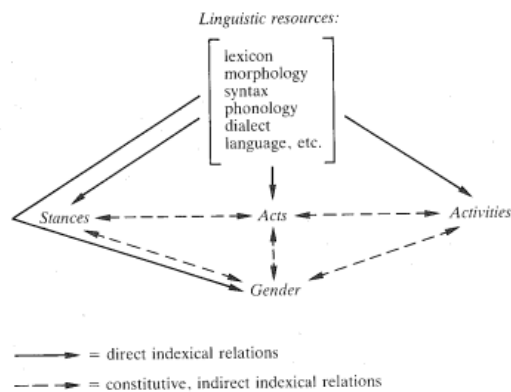


Figure 4: Direct and indirect indexes (Ochs 1992:342)

In Japan, for example, Ochs (1992:341) reports, the sentence-final particle *wa* directly indexes “delicate intensity”. Since being delicate is considered female in Japan, *wa* has become an indirect index of femaleness. Maleness, on the other hand, is indirectly indexed by the particle *ze*, which directly indexes “coarse”.

This notion of indexicality has been expanded, for instance by adding the notions of enregisterment (Agha 2003), orders of indexicality (Silverstein 2003) – how features acquire new meanings – and the concept of an indexical field (Eckert 2008b) which contains the potential meanings of a feature.

The shift in sociolinguistic methodology towards the social-constructivist model outlined in the last few paragraphs was later termed the “third wave of variation studies” (Eckert 2012a). Its impact on language and gender research is summarized by Cameron:

Whereas sociolinguistics traditionally assumes that people talk the way they do because of who they (already) are, the postmodernist approach suggests that people are who they are because of (among other things) the way they talk. This shifts the focus away from simple cataloguing of differences between men and women to a subtler and more complex *inquiry into how people use linguistic resources to produce gender differentiation*. It also obliges us to attend to the ‘rigid regulatory’ frame within which people must make their choices – the norms that define what kinds of language are possible, intelligible and appropriate resources for performing masculinity or femininity.

(Cameron 1997:49)

2.4.3. Recent Research

This social-constructivist approach inspired a range of rather diverse variationist studies. Among the most influential are papers by Bucholtz (1996; 1999b; 1999a), Kiesling (1998; 2007; 2009), Hall (1995), Hall & O’Donovan (1996), and Eckert (2000). Each is briefly summarized below.

Bucholtz (1996) shows how female “nerds” construct their own version of femininity in an American high school. Bucholtz argues that they do so by distancing

themselves from the “cool” girls and by developing nerd-specific styles. Difference to the cool girls is linguistically marked by speaking in a lower pitch than cool girls do and avoiding slang terms popular among the cool students. Nerd girls also participate less in an ongoing vowel shift. Regarding non-linguistic style, nerd girls prefer baggy and darker-colored clothes to the cool girl’s tight-fitting apparel. Bucholtz argues that the nerd culture places high value on being smart (rather than cool). Thus, nerd girls often speak in a more learned register (For instance asking *Is anyone knowledgeable about X?*), they enunciate clearly and avoid features of casual speech such as consonant cluster reduction. Bucholtz argues that nerd girls’ pride in their intelligence actually requires them to construct their own version of femininity different from the hegemonic femininity (represented by the cool girls) that does not allow for female intelligence. Nerd girls are only allowed to be smart because they reject traditional femininity (Bucholtz 1996:124). Bucholtz also points out that there is an ethnic aspect to her study: the nerd is very much a “white” category.

Another instance of gender and ethnic identity interacting is presented in Bucholtz (1999a). Analyzing his narrative of a fight, Bucholtz discusses how a white teenager uses linguistics means to first establish ideologies of masculinity and then align himself with one of them. She argues that the speaker invokes three ideological connections in his narrative: a gender ideology that links masculinity to physical power; a racial ideology that links blackness to physical power; and a language ideology that links African-American Vernacular English (AAVE) to both blackness and masculinity. In her data, this plays out as follows: narrating the fight, the speaker constructs an African-American version of masculinity that is characterized by physical power and violence, contrasting it with his own white masculinity that falls short in these regards. He does so by quoting his African-American opponent, who is speaking African-American

English and who questions the speaker's masculinity by calling him a *pussy* and a *punk*. Later on in his narrative, the (white) speaker claims this African-American physical masculinity for himself by inserting AAVE-features into his own language, for instance glottalizing the word-final /d/ in *dude*. "Narrative choices, including language crossing [into AAVE]", Bucholtz (1999a:455) summarizes, "allow [the speaker] to borrow an honorary black status and its accompanying ideological form of masculinity as developed earlier in this narrative".

Kiesling (1998; 2007; 2009) studied the construction of masculinities in a Virginia fraternity. Ultimately, he writes, every linguistic performance by a man is evaluated in relation to four discourses of masculinity in the U.S.: gender difference, heterosexism, power, and male solidarity (Kiesling 2009:197). In his study, he focuses on the power variable by quantifying the realization of (ing) as [ɪn] or [ɪŋ] among the fraternity brothers in various settings. In meetings – as opposed to “just hanging out” – most of them decrease the frequency of the vernacular variant [ɪn]. A few speakers, however, do not: Kiesling (1998:69) argues that they “use (ING) to index working-class cultural models and confrontational stances, as part of identity displays based on physical, rather than structural power”. This contrasts with the [ɪŋ]-using fraternity members, who rely on structural power: their use of the “correct” pronunciation connects them to qualities associated with people in positions of structural power in society, in the process creating power for themselves. These different kinds of power, Kiesling argues, are central to ideologies of masculinities in the U.S. He also points out that the linguistic indexing is usually combined with other stance-taking activities such as sitting in a certain corner of the room.

Hall (1995) analyzed the language of phone sex workers performing the “the ideal women” (Hall 1995:190) for their male clients over the telephone. Hall (1995:199–200)

found that performers engaged in what she calls “discursive shifting”: they shifted into a higher pitch and exaggerated or introduced discourse features such as questions and supportive comments. Some phone-workers in Hall’s study also reported using more intensifiers, more color terms, and a dynamic intonation pattern when performing on the phone. That way, Hall (1995:201) argues, the speakers –among them a male performer – “produc[e] a language that adheres to a popular male perception of what women’s speech should be: flowery, inviting, and supportive”.

Similarly, in their study of language use among the hijras, groups of male-born Indians who are usually considered neither men or women, Hall & O’Donovan (1996) analyze use of grammatically gendered forms. Hall & O’Donovan (1996:245) find that hijras switch between forms depending on the gender identity they are performing at that moment and depending on the gender of their addressee. These linguistic forms are also used to signal distance or solidarity; for instance, addressing another hijra with male-gendered forms is usually intended as an insult (Hall & O’Donovan 1996:251).

In a study of fraternity men, Cameron (1997) analyzed a casual conversation to investigate how the speakers construct their masculinity by distancing themselves from another man they describe as “gay” since he does not conform to their standards of masculinity. Cameron notes that they define their own masculinity by comparing it to this specific “antithesis of man” (Cameron 1997:59). By discussing this topic, they also defuse the danger of their all-men-hangout being considered a homo-erotic, rather than homo-social, event. Not being gay is, as indicated before, important to their concept of masculinity. Cameron notes that during their conversation, which could be categorized as “gossiping”, the young men engage in the joint production of discourse and share they linguistic floor – hallmarks of a speech style often considered feminine. Cameron argues

that concepts like “feminine” or “masculine” style cannot be established independent of the specific context.

Eckert (2000) analyzes gender variation in a study of high school students in the Northern U.S. (aspects of this research have already been discussed in section 2.4.1. above). Eckert (2000:55–8) describes how gender relates to differences in group affiliation (Jocks and Burnouts) by analyzing clothing style, hangout spots and drug use. Her data on vowel realization among the teenagers show an interaction between gender, jock-burnout status and urban-suburban orientation. The urban-linked variables, she argues, function as symbolic resources associated with local and institutional practice. Boys show more engagement with the sub-urban variables, while girls are especially intricate in their use of urban variables. Eckert (2000:169) hypothesizes that this is a reaction to gender ideologies: the use of urban variants is a threat to a girl’s purity while sub-urban forms may cast doubt on a boy’s toughness. “Gender groups, in other words, show more delicate use of variables that pose a greater potential threat to standard gender norms” (Eckert 2000:170).

As the studies cited above show, more recent variationist studies on gender share certain characteristics that influence their findings. In summary, these include:

- 1) A social-constructivist approach that treats gender as a performance. This entails increased agency on the speaker’s part: language choices no longer passively reflect an identity, but actively create it. This approach also replaces the binary gender model (male versus female) with a dynamic model that allows for a variety of masculinities, femininities, as well as identities defying all these categories.

- 2) The studies are quite diverse in their approach and methodologies. Bucholtz's work, for instance, is strongly influenced by conversation analysis while e.g. Kiesling uses essentially Labovian methods.
- 3) These studies are small-scale; the research cited above analyzes data from one (Bucholtz 1999a) to eleven (Kiesling 1998) speakers. This is partly a result of substituting an ethnographic approach for the stratificational model.
- 4) Linguistic forms are no longer ascribed prestige; rather, they are thought to index stances or qualities.
- 5) The focus has shifted from entire vowel systems or vowel shifts to individual features thought to be indexical.
- 6) Often, researchers take other stylistic features such as clothing into account when analyzing their findings.
- 7) Crucially, gender is related to or analyzed in relation to other social factors. Some studies above, for instance, find ethnicity to be an important factor: in Buchholtz's (1999a) study, black masculinity is different from white masculinity.

Another category that very prominently interacts with gender and that is also relevant to the study at hand is sexual orientation (Bucholtz & Hall 2004; Eckert 2012b). "Heterosexism", the assumption that men are interested in female partners, for instance, is part of Kiesling's (1998:74) concept of hegemonic masculinity in the United States and is implicit in Eckert's study (2000). It also informs the homophobic discourse analyzed in Cameron (1997) and the "sexy" feminine performances in Hall (1995).

This increased interest in analyzing gender alongside sexuality is reflected in the relevant handbooks as well. In its second edition, *The Handbook of Language and Gender*

(Holmes & Meyerhoff 2003), for instance, was renamed to *The Handbook of Language, Gender and Sexuality* (Ehrlich, Meyerhoff & Holmes 2014), in order to “highlight the ongoing importance of sexuality to the field and the close connections between gender and sexuality” as the editors write in the introduction (Ehrlich & Meyerhoff 2014:1). In the most recent edition of the *Handbook of Language Variation and Change*, the chapter on “Sex and gender in variationist research” (Cheshire 2002) was replaced by a text on “Gender, sex, sexuality and sexual identities” (Queen 2013).

2.5. LAKOFFIAN LANGUAGE AND GENDER STUDIES

The development within quantitative sociolinguistics mirrors a broader trend within linguistics. In “language and gender” studies, one can observe similar trend. The line between Lakoffian “language and gender” studies and the Labovian variationist, quantitative studies outlined above is not clear-cut. Language and gender studies here refers to research inspired by Lakoff’s study (1973; 2004) of “women’s language”. This research agenda differs from Labovian sociolinguistics mainly in its focus on gender as the only explanatory variable; less emphasis on empiricism and a feminist agenda.

Lakoff’s (1973; 2004) developed the idea of a female register, a “woman’s language”. According to Lakoff (1973:50–2), this register is characterized by the use of color words such as *mauve*, “meaningless” particles such as *oh dear*, and “female adjectives” like *divine* or *charming*. Women are also more likely to use tag questions. Lakoff argues that these linguistic characteristics conspire to make women seem timid, even without an opinion. Lakoff (1973:76) then relates these linguistic characteristics to the broader social context: “the social discrepancy in the positions of men and women in our society is reflected in linguistic disparities”. Lakoff’s idea that language reflected the social

oppression of women was later dubbed the “dominance approach” to gender variation, to distinguish it from the competing “difference approach” (sketched in section 2.5.2. below).

2.5.1. The dominance model

This “dominance approach” was further pursued by studies analyzing use of the features identified by Lakoff and relating them to male dominance.

Fishman (1983), for instance, found women to be the conversational “shitworkers” in interactions of heterosexual couples: women worked hard to initiate and maintain conversations, while their male interlocutors did not pull their own weight. Similarly, in their study of conversations between men and women, West & Zimmerman (1983) found men to interrupt women more than the other way around. They interpreted this as reflecting the male-empowering social system. O’Barr and Atkins’ (1980), analyzing court hearing, found that what Lakoff terms “women’s language” overlapped to a large extent with features used by speakers in positions of powerlessness points in the same direction. (Their study looked at language use in court hearings).

The findings of studies in the Lakoffian dominance paradigm were later challenged by Cameron et al (1988) and Holmes (1990), who argued that the features of “women’s language” need not always be indicators of powerlessness. Holmes (1990) argues that Lakoff oversimplified the meaning and function of tag questions like *you know*: contrary to Lakoff’s claims, they do not always indicate weakness or inferiority; and they are not used more frequently by women than by men in general, according to Holmes. *You know*, for instance, was used twice as frequently by the men in Holmes’ corpus (Holmes 1990:199). Looking at the semantics of *you know*, Holmes identified two different functions: it can indeed be used to either express uncertainty in the sense of Lakoff but it can also signal confidence or certainty (e.g. *I’m the boss around here, you know*: Holmes

(1990:189)). Holmes found that, contrary to Lakoff's thesis, women used less *you knows* overall, and actually used it more than men in the confidence-indicating function introduced above. Cameron summarizes her and Holmes' critique of Lakoff's approach as follows:

The relation between linguistic form and communicative function is not a simple thing and we cannot state *a priori* what tag questions do [...] This should make future researchers wary of the line of argument popularised by Lakoff, that if women use form x more than men we should seek an explanation of this in terms of the invariant communicative function of x.

Cameron (1988:91)

Elsewhere, Cameron (1992a:24) criticizes the dominance approach as "reductive" in that it disregards so many aspects of gender relations other than power – she mentions desire and sexuality.

2.5.2. The difference model

Other researchers unconvinced by Lakoff's focus on male dominance while still trying to explain linguistic gender differences developed what was dubbed the "difference" approach: instead of appealing to social dominance as the explanatory factor, they argued that men and women form two distinct linguistic cultures. First formulated by Maltz & Borker (1982), the argument is that men and women, who typically play in gender-segregated groups during childhood, belong to different sociolinguistic subcultures with different conversational styles.

Goodwin (1980), for instance, established that boys playing in hierarchically organized groups develop a competitive conversation style while girls, who tend to play in smaller egalitarian groups, a co-operative conversation style. These differences are

thought to persist into adulthood and lead to different speech patterns. This conception was introduced to a wider audience by Tannen's (1990) best seller *You just don't understand*. Feminist researchers saw this line of argument as an opportunity to address the negative portrayal of women's language as timid and powerless by Lakoff. Studies by Holmes (1993) or Coates (1989), for instance, attempted to show the benefits of the female conversational style. In her study of New Zealand English, Holmes argues that their frequent use of pragmatic particles, their abstaining from interruptions, and their use of compliments and apologies makes women "ideal conversational partners" (Holmes 1993:91). Coates (1989:120) writes that in her study of all-female interactions, "the way women negotiate talk symbolizes mutual support and co-operation".

This difference approach has been criticized (Cameron 1992a) for a static understanding of language: in this paradigm, norms picked up during childhood determine lifelong linguistic behavior. Arguing for the importance of power, Cameron (1992a) notes that the development in gender-segregated groups might be affected by social power structures (or that the power structures lead to gender-segregated play groups in the first place). Some of the criticism leveled at Tannen (1990) applies to the paradigm in general: the book's argument has been criticized as de-political and even anti-feminist by feminist researchers (Troemel-Ploetz 1991; Cameron 1992b; Freed 1992) who see it to justify or ignore social inequalities in gender relations.

2.5.3. Linguistic representation of women

Besides the research on differences in language use, Lakoff also investigated the linguistic representation of women: how do we speak about women and what does it tell us about social structures? Lakoff (1973:63) shows that words associated with women tend to acquire negative connotations: compare for instance male *master* to female *mistress*.

Lakoff (1973:58–61) also discusses a tendency towards describing women in relation to their marital status, most explicitly in terms of address such as *Mrs John Smith*; she investigates the use of *lady* as an euphemism for *woman*. Research following in this strand of the Lakoffian tradition has often focused on job titles, more specifically the use of generic masculine terms (*chairman*) and linguistic ways of turning women into a marked category (*doctor* versus *lady doctor*). Terms of address (*Mrs.* versus *Ms.*) were also researched extensively. An important goal of this research strand was advocating for non-sexist language use. This line of research will not be discussed in more detail here, as it is not relevant to the study at hand. Detailed overviews can be found in Eckert & McConnell-Ginet (2013:193–222) and in Romaine (1999:91–149). Some influential papers are compiled in Cameron (1998a:83–164).

2.6. GENDER VARIATION IN WRITING

All the research cited above looked at spoken language exclusively. Gender variation in writing, which is the subject of the present study, has not been studied as extensively (Lillis 2013). Studies by Koppel et al (2002), Mulac & Lundell (1994) and Newman et al (2008) are notable exceptions. Their papers, however, differ remarkably in methodology and results. (Studies on writing in computer-mediated communication (CMC), which are even more relevant to the present study are discussed in more detail below).

Newman et al (2008), for instance, studied gender differences in a corpus of more than 45 million words compiled from various sources. The results relevant to this study are that women tend to use more pronouns and verbs, while men commonly use longer words and more articles and numbers. Mulac & Lundell (1994) compared essays written by female and male students; among their findings is a tendency among men to use more

numbers while female writers are more likely to use progressive verbs and write longer sentences. Koppel et al (2002) designed a text classifier that was able to quite reliably group texts from the British National Corpus according to author gender. The features their algorithm made use of included noun specifiers (determiners, numbers) as an indicator of male writing and pronouns as an indicator of female writing.

Previous studies on language and gender in *online* writing (CMC) – as opposed to writing in general – include studies of blogs by Herring and Paolillo (2006), Nowson et al. (2005) and Schler et al. (2006). There is also a growing body of research on language use on Twitter, such as Rao et al. (2010) and Bamman et al. (2014). This research is summarized below.

In their study of a 35,000 word corpus of blog posts, Herring & Paolillo find that one cannot study linguistic variation online without taking the context – and content – of the blog into account. In their dataset, women for instance used more personal pronouns. This is not because of a gendered style, Herring & Paolillo argue, but because more women write online diaries, a genre that requires the use of pronouns. Herring & Paolillo conclude that:

The results show that the diary entries contained more ‘female’ stylistic features, and the filter entries [on blogs commenting on politics and business] more ‘male’ stylistic features, independent of author gender. These findings problematize the characterization of the stylistic features as gendered, and suggest a need for more fine-grained genre analysis in CMC research.

Herring & Paolillo (2006:439)

In a similar study of a 300 million word corpus of English blogs from *blogger.com*, Schler et al. (2006) find words in the lexical groups “money”, “job”, “sports”, and “TV” to correlate with male authors. Female writers, on the other hand, tended to use more words

from the “sleep”, “food”, “sex” and “family” lexical fields. In addition to these content words, Schler et al. also studied what they call stylistic features: Schler et al. find that women tend to use more pronouns and neologisms such as *lol* as well as assent or negation words; men tend to use more articles and prepositions. Schler et al. argue that this indicates that women favor an involved, men an informational writing style.

In their study of weblogs (410,000 words), Nowson et al. (2005) place blog writing on a formality-scale. Using a classification scheme developed by Heylighen & Dewaele (2002), they find blog writing to be less formal than email but more formal than biographies from the British National Corpus.

In addition to blogs, Twitter has emerged as one of the premier online research venues. Rao et al. (2010), for instance, work with Twitter data to design a tweet classifier with 70 percent accuracy in assigning gender. Some of the most important features the algorithm makes use of are emoticons and the alphabetism *OMG*.

Bamman et al. (2014) study variation in a corpus of 9 million tweets from 14,000 Twitter users. They attempt to show that the traditional binary gender concept is not sufficient to explain the variation found in their dataset, arguing that “[i]f we start with the assumption that ‘female’ and ‘male’ are the relevant categories, then our analyses are incapable of revealing violations of this assumption” (Bamman, Eisenstein & Schnoebelen 2014:148). Their study thus first identifies gender markers in the traditional sense by comparing usage frequencies between women and men, finding that women used more pronouns, emotion terms and “CMC words” such as *lol* or *omg*. Men, on the other hand, tended to use more numbers, technology words, and swear words.

Bamman et al. then cluster all their speakers according to their usage frequencies of the 10,000 most common lexical items. This way, they discover topic and style-defined clusters that are often dominated by one gender. Male-dominated clusters, for example,

include ones centered around topics like sports or computer programming. Finally, Bamman et al look at speakers who do not follow the statistical trends for their gender group. They find that these speakers still follow the norms for the context they engage in: their social networks are typically dominated by members of the gender group whose linguistic style they are accommodating to.

In addition to the academic research presented above, several popular science books taking a large-scale quantitative, “big data” approach to gender and language variation, have recently been published. The most relevant to the present study are Rudder’s (2014), which discusses, among other aspects of online dating such as age preferences, linguistic behavior on the dating website *OkCupid*, and Blatt’s (2017) study of language use in fiction writing. Rudder (2014), in his monograph *Dataclysm*, identifies words preferred or disfavored by ethnicity and gender. For instance, according to his analysis of dating profiles, the “most typical words” for white men include *blue eyes* and *Allman Brothers*; for black men, *dreads* and *Jamie Foxx* are among the most typical. On the other hand, Latina women disfavor the word *Cincinnati*, black women the verb *tanning* (Rudder 2014, Part 3.10). Regarding the genre of dating ads, however, Rudder (2014, Part 3.10) argues that it does not constitute a productive environment to study gender variation in: word-based analysis, as the ones outlined above, he argues, will just identify the addressee as the most distinctive lexical item (for example, heterosexual women are distinguished from the rest of the dataset by using more words describing and referring to men). Rudder (2014, Part 3.10) writes that sex and profile text are “inextricable” to an extent that makes analysis of linguistic gender differences effectively impossible.

The second entry in this list of recent “big data” language publications is Blatt’s (2017) *Nabokov’s favorite word is mauve* (incidentally, the word *mauve* is one of the features of Lakoff’s women’s talk). Blatt (2017) presents a computational text analysis of the

literary canon. In the chapter on gender, Blatt (2017, chapter 2) identifies words indicative of male and female author gender (such as *chief*, *rear*, *civil* and *pillow*, *lock*, *curls* respectively, in classic works of literature: Blatt (2017, chapter 2)), ranks book by masculinity and femininity of word choice, and computes *he-she* ratios to detect gender bias in works of fiction. Blatt's (2017) research differs from more academic linguistics studies in that he does not attempt to control for non-linguistic variables such as genre or time period. In addition, he does not pay much attention to an item's linguistic context, for instance distinguishing the author's voice from a character's voice when studying author style.

2.7. IMPLICATIONS FOR THIS STUDY

This, then, is the context that the present study is situated in. Based on the previous research and discussions within the field outlined above, the following approach is taken to the study of gender variation.

First, a dataset that foregrounds and affords gendered performances, that is online dating ads, is chosen and maximized for sample size. A feature set appropriate to this genre, namely features frequent in computer-mediated communication, is used. Relevant social groups and linguistic features are not presupposed, but as much as possible established through empirical analysis.

Identification of gendered forms is achieved by combining an analysis of linguistic production with an analysis of language perception, requiring that they must exhibit consistent results regarding the feature's social relevance. A thorough analysis of social and linguistic context then informs the argument about which indexical meaning is activated in the given context. This analysis relies on empirical findings regarding text genre, gender ideologies, and audience effects, combining external research with findings from the perception study.

Below, chapter 3 presents a study of language production in dating ads. Chapter 4 discusses the study of language perception, before results and implications are discussed in chapter 5.

Chapter 3: Production

This chapter presents the first part of the linguistic analysis, a study of language production. A perception study based on its findings is discussed in chapter 3. Chapter 4 will discuss the insight gained from both studies.

3. 1. THE RESEARCH QUESTION

The overarching question of the whole study is to investigate how people use linguistic resources to produce gender differentiation. As noted before, the process of gender differentiation involves two parties: the speaker and their audience (Agha 2007; Eckert 2008a). The audience's side of the equation, that is: gender *perception*, will be addressed in chapter 3. In this chapter, the focus is on the *performance* of gender. Analyzing a set of dating ads, we ask: is there gendered language variation in these ads? If so, what linguistic features are pertinent to gender? And what aspect of gender does this variation depend on – the author's gender, the author's sexual orientation, the addressee's gender, the addressee's sexual orientation?

Thus, the hypotheses to be tested are:

- H1 There is variation by binary author gender (male/female) in the data.
- H2 There is variation by binary addressee gender (male/female) in the data.
- H3 There is variation by the author's and addressee's sexual orientation (binary: heterosexual/homosexual) in the data.
- H4 There is variation by an interaction of the above in the data.
- H5 The variation reflects American gender ideologies.

To empirically address these hypotheses, a dataset of 103,290 personal ads from www.craigslist.org was compiled, each containing information about the author's gender and their addressee's gender (see section 3.2.). A set of 9 linguistic features including

abbreviations, emoticons, and non-Standard punctuation is extracted from each ad (section 3.3.1.). These features are then used to cluster the ads by feature usage (section 3.3.2.). The resulting clusterings will allow us to speak to points H1 – H3, namely whether author gender, addressee gender, or sexual orientation are most relevant predictors of linguistic variation in the data. Based on the results of the clustering analysis, a more fine-grained statistical analysis and visualization of results will be conducted (section 3.3.3.). First, an overview of previous research with dating ads within various scholarly disciplines is given in section 3.1.1.

3.1.1. PREVIOUS RESEARCH

Earlier research across several disciplines used personal ads as data in the context of the so-called social exchange theory (Lance 1998:299): in this framework, dating ads are analyzed as an interaction of two parties looking to strike a deal, a social exchange. The writer, according to the theory, offers certain social goods, looking for a partner's social goods in return. By counting the occurrence of words from semantic categories such as “physical characteristics” or “professional attainment”, researchers try to evaluate what attributes are perceived as particularly valuable for each gender. For instance, if men tend to focus on words from the “professional attainment” category in their self-description, this might indicate that these attributes are seen as a valuable good to offer a prospective mate. If, on the other hand, their description of the desired female partner focuses on physical characteristics, these might be the social goods valued in women. Economists used this approach to study self-marketing and decision making, sociologists to gain insight into gender role expectations, and linguists to study linguistic gender ideologies. A brief overview of their findings is given below.

Economists and marketing researchers have used dating ads in studies such as Peters et al. (2013), who studied online dating ads as the “ultimate form of self-marketing” (Peters, Thomas & Morris 2013:80). Working with 1,200 ads from Craigslist, they identified various styles of self-marketing, the most pronounced of which they called the “sales pitch”, mainly employed by male writers. Economists Hitsch et al (2010) analyzed the word use of 6,485 online daters from the perspective of preference theory. Their data indicates that writers are usually searching for someone similar to themselves, especially when it comes to age and ethnicity. Regarding gender differences, they note that women seem to be more influenced by income, men more by looks.

Sociologists have used dating ads to study gender roles in American and British society. Typically, sociological studies work with several hundred ads they code for semantic categories as described above: Davis (1990:47) for instance, computed how often writers mention physical attractiveness, occupation, and financial security in 329 ads. The social exchange facilitated in dating ads, according to Davis’ results, can be described as beauty for money: men look for youth and physical attractiveness, offering financial security and social status; women conversely ask for financial success and social status in their partner, offering youth and physical attractiveness. In Davis’ study, this is reflected by the fact that men were more likely to present their social status or career path in their self-description; women tended mention their looks and age. (Davis’ study summarizes the idea in its title: “Men as success objects and women as sex objects”). The same dynamic is illustrated by the findings of Lance (1998), who looks at 1,400 newspaper ads, and Koestner & Wheeler (1988), working with 400 ads. Koestner & Wheeler note that social exchange theory can be extended to specific physical characteristics, finding that women look for height and offer (low) weight, while men do the reverse. Smith & Stillman (2002) expand on the theme further by including homosexual writers. They argue that sexual

orientation is an important predictor of ad content as the writers take the (assumed) desires of different audiences into account when setting up the social exchange deal. Coming from a slightly different theoretical perspective, Jagger (1998) argues in her study of postmodern dating patterns that lifestyles choices (hobbies, consumption patterns) have replaced status and physical criteria as differentiators in dating ads. The same point is made by Fullick (2013), working with 20 profiles from *nerve.com*. Additionally, Jagger (1998) argues that in a postmodern dating environment, both genders have to offer an attractive body; “men too have become tyrannized by the aesthetical ideal of ‘slenderness’” (Jagger 1998:806).

Linguists – mostly discourse analysts – have used dating ads to study identity formation. Notable examples include Coupland (1996; 2000), Thorne & Coupland (1998), Winn & Rubin (2001), Marley (2007; 2008a; 2008b), Ellison et al. (2006), and Fullick (2013). They all note that the dating ad genre foregrounds the construction of a persona through language alone, as a “distilled social identity performance” (Winn & Rubin 2001:394) or “rich and dense articulations of identity” (Thorne & Coupland 1998:234). They argue that studying dating ads as such vehicles for self-presentation can give insight into the construction of linguistically gendered characters. Besides this socio-linguistic aspect, these studies are primarily interested in what Coupland (1996; 2000) calls the commodification of the self: how writers are turning themselves into a “marketable product” (Fullick 2013:555). This act of identity creation happens in a very formulaic genre, reminiscent of used car ads or job postings (Coupland 1996:188; Fullick 2013:548), making analysis and comparison easier. Coupland (2000:10) formalizes dating ads’ usual structure as follows:

(a) advertiser (b) “seeks” (c) target (d) goals (e) comment (f)
reference.

Within this structure, ad writers establish certain identities through both the topics they choose to bring up and the linguistic features they choose to use.

The first aspect – what content words are used by writers? – is described for heterosexual writers in Coupland (1996), who finds a focus on age, appearance, and personality in 100 newspaper ads. Thorne and Coupland (1998) take the same approach to 200 ads posted by homosexual writers and find pervasive self-masculinization and a strong focus on the body among male gay writers, a discourse which has no equivalent among the lesbian writers. Rubin and Winn (2001), on the other hand, focus on linguistic features such as commas or pronouns use. In their study (they have 84 participants write and respond to dating ads) they find that female writers tend to use more markers of non-essential information (e.g. dashes) and excitability (e.g. exclamation marks) than male writers (Winn & Rubin 2001:309). Ellison et al.'s (2006) study of 36 online daters, while not focusing on any specific linguistic items, concludes that “[s]tylistic aspects such as timing, length and grammar appear equally important as the content of the message” (Ellison, Heino & Gibbs 2006:431).

All the studies cited above also complicated initial notions about which linguistic features are relevant to gender. They all point to the overarching relevance of genre as an external variable that trumps gender, Winn & Rubin (2001) and Rubin & Greene (1992) most emphatically. They, see for instance Rubin & Greene (1992:399) and Thorne & Coupland (1998:246), also stress that gender boundaries are not clear-cut, and a one-on-one mapping of gender and linguistic features is unrealistic. This echoes theoretical points discussed in the language and gender research tradition, discussed in chapter 2 (for instance Ochs (1992)).

3. 2. THE DATASET

The present study, while to some extent trying to address the same issues as the papers cited above, differs in regard to the size of the dataset – several thousand rather than several hundred ads – as well its methodology, a computational text analysis rather than manual counting of words. The dataset is introduced below, the methodology is outlined in section 3.3..

The analysis presented in this chapter is based on a corpus of 103,290 personal ads (10,065,808 words) downloaded from the American section of www.craigslist.org. Craigslist is a website for local classifieds, active in more than 700 locations in 70 countries (Craigslist.org 2016). It currently attracts more than 60 million unique visitors per month (Bensinger 2017). The website offers a “Dating” section, where ads can be filed under the labels “w4w” (women looking for women), “w4m” (women looking for men), “m4w” (men looking for women), and “m4m” (men looking for men). It also offers a “Strictly Platonic” section for friendship or companionship-related ads that is categorized in the same way. All ads found in these sections were downloaded in the summer of 2015.

To prepare the data for analysis, Python scripts were used to delete Spanish-language posts and delete duplicate postings. (All the scripts used are available online on www.github.com/patrickschu/chapter2). Spanish-only posts were identified by comparing a ten-word-chunk from each ad to the words in the eow Celex-database of English (Baayen, Piepenbrock & van Rijn 1993). Ads containing chunks with a match ratio of less than 3 were removed automatically; ratios of 4 – 6 were deleted only after inspection of the entire ad. To identify duplicates, all files with matching titles and same length were automatically deleted. If two files shared the exact same ten first words, one was deleted. The final corpus consisted of 167,180 ads, 103,290 of them dating ads (called the “dating ad corpus” from here on) and 63,890 “Strictly Platonic” ads (called the

“Strictly Platonic ad corpus” from here on). The dating ad corpus contains 10,065,808 words, the Strictly Platonic ad corpus contains 5,462,614 words for a total of 15,528,422 words. Only the dating ad corpus was used for this study. A breakdown of the number of ads and words in the dating corpus by category is presented in table 1 below. Categories that apply to fewer than 50 files are not listed, but included in the sums. Labels not included in the four Craigslist-categories listed above, such as *mw4mw* occur when users change the ad’s headline to fit their own needs, in this case a male-female couple looking for the same.

Gender category	Ads	Words
m	74,889	7,457,383
w	27,994	3,132,928
mw	276	25,985
t	56	4,795
mm	52	3,808
ww	17	1,507
wm	6	620
Total	103,290	10,627,026

Table 1: Number of ads and words by gender in the dating ad corpus.

Category	Number of ads	Number of words
m4w	48,772	5,640,390
m4m	25,671	1,785,778
w4m	15,769	1,831,117
w4w	12,169	1,297,160
mw4w	223	21,688
m4mm	165	12,167
m4mw	159	11,266
m4t	77	4,310
Total	103,290	10,627,026

Table 2: Number of ads and words by category in the dating ad corpus.

Addressee	Number of ads	Number of words
w	61,195	6,962,734
m	41,525	3,623,462
mw	214	14,903
mm	187	14,218
ww	76	6,325
t	87	4,869

Table 3: Number of ads and words by addressee gender in the dating ad corpus.

Note that the tables above only refer to the dating ad corpus.

This dataset allows us to explore some avenues previously not investigated in much detail in the language and gender research.

- 1) While gender variation is widely studied in spoken language, variation in writing has not been given the same amount of attention (Lillis, Davis?).
- 2) In addition to variation by the author's gender, this dataset allows us to take potential effects of addressee gender or the author's sexual orientation into account.
- 3) Genre effects are controlled for by selecting only one type of text, namely dating ads.
- 4) All participants are "self-categorized"; authors, rather than the researcher, determine what gender group they consider themselves part of by posting under the respective label.
- 5) The dataset is larger than most comparable studies, increasing the statistical power of the study.
- 6) Selection of relevant linguistic is part of the study. As few assumptions as possible are made pre-analysis about what linguistic features are gendered.

At the same time, this dataset suffers from various weaknesses.

- 1) There is no information about the author besides the text. For this reason, we are not able to control for – or correlate with – age, social class or other variables commonly investigated in sociolinguistics.
- 2) In self-labeled data, writers can be deceptive about gender, audience, or sexual orientation.
- 3) The dataset is too big for an exhaustive qualitative analysis of the data. Instead, individual ads will be picked out to illustrate features.

With this in mind, a discussion of methodology and findings is presented below.

3.3. METHOD

This section provides an introduction to and justification of the research methodology adopted to address the hypotheses outlined above (reproduced here for convenience).

- H1 There is variation by binary author gender (male/female) in the data.
- H2 There is variation by binary addressee gender (male/female) in the data.
- H3 There is variation by the author's and addressee's sexual orientation (binary: heterosexual/homosexual) in the data.
- H4 There is variation by an interaction of the above in the data.
- H5 The variation reflects American gender ideologies.

To address these hypotheses, the study progresses in two steps. First, it needs to establish whether there is linguistic variation in this dataset and if so, which features are relevant (“*discovery*”). Second, it needs to describe and analyze existing patterns (“*analysis*”). Third, it will present results.

The process of discovery, described in section 2.3.2, in this study consists of clustering the data based on linguistic features. The analysis will build on the results of the clustering analysis, by conducting a statistical analysis including hypothesis testing and visualization of patterns. This process is presented in the sections below as follows: first, features and feature extraction (2.3.1) to prepare the data for clustering; second, data clustering (2.3.2); third, a statistical analysis of patterns suggested by the clustering (2.3.3); and finally the presentation of results.

3.3.1. Feature extraction

The starting point for the identification of linguistic features will be the concept of *e-grammar* (Herring 2012). The e-grammar model is an attempt at a comprehensive compilation of features of computer-mediated communication by grammatical category. The model consists of three levels: the *microlevel* contains typographical and orthographical features; the *word level* roughly corresponds to morphology; the *utterance level* addresses features on the sentence level. In the following, the attempt at extracting the features contained in the e-grammar mode will be outlined. Within the time and financial constraints of the study, only the micro- and the word level will be covered in the following. (Note that the order of features in this study does not exactly following Herring's; in the present study, features will be ordered alphabetically, mainly to aid with reading of the visualizations to follow).

Levels	Features
The microlevel	
Typography	1) Emoticons (e.g. ;)) 2) Repeated punctuation (e.g. <i>Really?!?</i> , Squires 2012) 3) Replacing words with numbers (e.g. <i>This is 4 you</i>) 4) Leetspeak, a kind of online jargon that is increasingly used in mainstream communication, according to Herring (2004)
Orthography	5) Abbreviations, acronyms (e.g. <i>lol</i>), clippings, vowel omission (e.g. <i>pls</i>) 6) phonetically motivated letter substitution (e.g. <i>s</i> → <i>z</i> in: <i>I can haz cheezburger</i>) 7) eye dialect (e.g. <i>sez</i>) 8) spellings that represent prosody or non-linguistic sounds such as laughter (e.g. <i>haha</i>)
The word level	
	9) some word formation processes such as acronyms seem more productive in CMD 10) neologisms (e.g. <i>newbie</i>) or conventionalized typos (e.g. <i>pron</i>) 11) specific leetspeak suffixes: <i>-zor</i> , <i>-zorz</i> , <i>-age</i> (Herring, 2012, p. 3)
The utterance level	
	12) elision, especially of articles and pronouns 13) double inflected modals (e.g. <i>I can haz</i>) 14) nominalization of verbal predicates (e.g. <i>Austin is the rocks-er</i> , meaning <i>Austin really rocks</i> ; cf. Herring, 2012, p. 4) 15) emotes (e.g. <i>*waves*</i>)

Table 4: Features of e-grammar, adapted from Herring (2012).

By focusing on e-grammar, this study does not make use of the feature set commonly associated with studies of language and gender. Features commonly studied include tag questions, hedges, and turn-taking (cf. Eckert & McConnell-Ginet 2013). While certainly relevant and worthy of study, several reasons make them not quite appropriate for this study. These relate to the dataset and methodology adopted here and the general research history of the traditional language and gender features.

First, most of the traditional features are more prevalent in speech than in writing. Turn-taking, for instance, is not possible in the written dataset used here. Other features such as question tags are exceedingly rare. In general, online writing calls for different feature sets than spoken language. The e-grammar model used here is one attempt to conceptualize such a feature set.

In addition to this, most of the traditional features require in-depth qualitative analysis. Holmes (1990) for instance shows that the meaning of tag questions needs to be determined on a case by case basis. This is why most studies including these features are now done in a discourse analysis framework – that is, taking a qualitative rather than quantitative approach – where this kind of attention to detail is possible.

In general, most of these features date back to the paper by Lakoff (1973) who suggests them as potentially gendered based on anecdotal evidence and personal observation. While this is certainly a valid starting point, it must be noted that the mapping between language and social context is often unpredictable and complex (Ochs 1990) and relevant features might thus not be immediately obvious to the researcher. After decades of empirical studies testing Lakoff's ideas, there is still no clear consensus as to their meaning or validity. It is thus unclear whether they are indeed gendered language features or whether their prevalence in language and gender studies is the

results of what Wareing (cited in Cameron 1998b) called a “Hall of Mirrors”-effect: everybody studies these features because everyone else does and results are over-generalized or compared to very different studies.

The extraction of the nine e-grammar features is described in the sections 2.3.1.1. to 3.3.1.10. below. These are alphabetisms and acronyms (3.3.1.1.), capitalization (3.3.1.2.), clippings (3.3.1.3.), emoticons (3.3.1.4.), leetspeak (3.3.1.5.), rebus (3.3.1.6.), repeated punctuation (3.3.1.7.), prosody (3.3.1.8.), words replaced by letters (3.3.1.9.) and a lexical analysis described in section 3.3.1.10..

3.3.1.1. Feature 1: abbreviations

Following the definition in Chrystal (2011), *acronyms* and *alphabetisms* are types of abbreviations, both created by concatenating the initial letters of two or more words. Acronyms are pronounced as one word (e.g. *radar*) while alphabetisms are not (e.g. *CIA*). Both can be typed in all capital or all lower-case letters in English (Cannon 1989). In order to identify common alphabetisms and acronyms in the corpus, all words in capital letters that were labeled non-Standard by the automatic Python spell-checker PyEnchant (Kelly 2016) were inspected. This approach was chosen because these abbreviations are not accepted as words by a spellchecker and all-caps words are relatively rare even in this large corpus. Various misspellings and prosodic items such as *OHH* were discarded in the process. The meaning of each abbreviation was established by consulting online resources such as *Urban Dictionary* (Urban Dictionary 2017) and by inspecting the item in context, i.e. in the sentence extracted from the ad. The resulting abbreviations of interest were categorized by type, acronym versus alphabetism. They were then categorized as “noun” or “not-noun” to establish whether it could form a plural and the search algorithm would have to be adapted accordingly. Certain very frequent use cases such as state alphabetisms

(*TX, NY*) and school names (*NYU, KU*) were also labeled as such. Thus, for instance, the existence of *LTR* helped identify *ltr* as an alphabetism of *long-term relationship* in the corpus. It was then added to a list of three-letter abbreviations, labeled an alphabetism, and a noun. The resulting list of 284 abbreviations was integrated into a regular expression search on the untokenized corpus. State alphabetisms (which violate the rule that letters need to represent initial characters) and proper names such as schools or consumer brands (*BMW*) were excluded. Note that actual acronyms, that is abbreviations that can be pronounced as a word, are rare and pretty much limited to *YOLO* and potentially *LOL*.

Abbreviation	Tokens	Long version
<i>LOL</i>	6,068	<i>Laughing out loud</i>
<i>DDF</i>	5,859	<i>Drug and disease free</i>
<i>BBW</i>	5,113	<i>Big beautiful woman</i>
<i>OK</i>	4,624	<i>O.K.</i>
<i>FWB</i>	3,941	<i>Friends with benefits</i>
<i>LTR</i>	3,693	<i>Long-term relationship</i>
<i>HWP</i>	3,336	<i>Height-weight proportional</i>
<i>NSA</i>	1,891	<i>No strings attached</i>
<i>HIV</i>	1,575	<i>Human immunodeficiency virus</i>
<i>SWM</i>	1,547	<i>Single white male</i>
<i>HMU</i>	1,491	<i>Hit me up</i>

Table 5: Ten most common acronyms and alphabetisms in the dating corpus.

3.3.1.2. Feature 2: capitalization

Three patterns of non-Standard capitalization were found in the corpus: all-caps realizations of words (*IAMFROMNEWJERSEY*) as well as instances of so-called camel case (*iPhone*) and Pascal case (*PreTTY*). A regular expression was compiled for each of these patterns. Acronyms and alphabetisms (already identified in feature 1) in all caps were not excluded as the writer also had the option to use lower-case letters for these. Note that the “Instances” and “Tokens” counts in the table below do not add up as a stretch of capitalized characters might contain several capitalized words.

Capitalization pattern	Instances
All caps	7,681
Camel case	664
Pascal case	604

Table 6: Instances of non-Standard capitalization patterns in the dating corpus.

Word	Tokens
<i>NOT</i>	4,181
<i>YOU</i>	2,546
<i>BBW</i>	2,196
<i>DDF</i>	2,189
<i>AND</i>	2,040
<i>FWB</i>	1,766
<i>MEN</i>	1,717
<i>PLEASE</i>	1,509
<i>HWP</i>	1,428
<i>THE</i>	1,358

Table 7: Ten most commonly capitalized tokens in the dating corpus.

Word	Tokens
<i>KiK</i>	42
<i>FaceTime</i>	39
<i>ParTy</i>	38
<i>YouTube</i>	33
<i>PhD</i>	30
<i>iPhone</i>	27
<i>LoL</i>	26

Table 8: The most common items in camel and Pascal case in the dating corpus.

3.3.1.3. Feature 3: clippings

Clippings – a type of abbreviation where parts of the word are dropped to create a shorter lexical item – were identified by searching the lexicon of words used in the dating corpus. The corpus was tokenized using the NLTK function `word_tokenize` (NLTK Project 2015). Then, chunks of 2 to 5 characters from the beginning, middle and end of the word were successively compared to the rest of the lexicon. That is, if the lexicon contained the items *pic* and *picture*, the search algorithm would identify *pic* as a potential clipping of *picture*; it would also consider it a clipping of *topic*. The results were manually checked for accuracy and inspected in sentence context. Only items that can be considered valid clippings within this specific dataset were included, that is to say items where both the clipped as well as the full word were found. *Bra* for instance, while technically a clipping of *brassiere*, was not considered a clipping due to the fact that the word *brassiere* never occurred in the corpus. *Grad*, on the other hand, was retained as the

lexicon also contained the noun *graduate*. Overall, a list of 140 clippings were identified this way. Token counts are presented in table 8 below.

Clipping	Tokens	Long form
<i>pic</i>	42,297	<i>Picture</i>
<i>stat</i>	8,118	<i>Statistics</i>
<i>bi</i>	5,344	<i>Bisexual</i>
<i>ad</i>	5,153	<i>advertisement</i>
<i>info</i>	2,130	<i>information</i>
<i>sub</i>	1,883	<i>Submissive</i>
<i>fem</i>	1,633	<i>Female, Femme</i>
<i>masc</i>	1,590	<i>Masculine</i>
<i>dom</i>	1,413	<i>Dominant, Dominator/trix</i>
<i>site</i>	1,391	<i>website</i>

Table 9: Counts of clippings in the dating corpus.

3.3.1.4. Feature 4: emoticons

The list of emoticons, defined as graphical representations of facial expressions, was downloaded from the Wikipedia page “List of Emoticons” (Wikipedia 2014). Emoticons were categorized by type-based Western style (lying on the side, e.g. :) , Asian-style (upright, (+ +)) and non-character based Unicode emojis (🤔). Overall, 379 different emoticons of the various types were included in the script, and identified through regular expressions on untokenized text. The regular expressions were flexible regarding white space, thus :) and :) are considered equivalent.





Emoticon	Tokens
:)	6,284
;)	2,166
	145
=)	138
:(137
	137
:D	133
	122
:P	120
:p	63
:/	53
	42

Table 10: Counts of emoticons in the dating corpus.

3.3.1.5. Feature 5: leetspeak

Leetspeak is an online-only type of spelling variation that relies on creative replacement of letters with number characters (Herring 2012:2). Examples include *g00d* for *good*, where the character *o* is replaced by the number *0*. To identify instances of Leetspeak, all items in the corpus lexicon were compared to each other. Items that differed only in one of the relevant pairs (e.g. *3* for *e*, *1* for *i*) were retained after further inspection of context. The resulting list of search terms was 80 items long. As the tables below show, this feature is rather rare.

Leet item	Tokens
<i>to</i>	6
<i>yah00</i>	5
<i>s3nd</i>	2
<i>k1k</i>	2

Table 11: Counts of Leetspeak items in the dating corpus.

3.3.1.6. Feature 6: replacing letters with numbers (rebus)

What is referred to as rebus here is a phenomenon similar to Leetspeak. However, while Leetspeak replaces individual *characters* with numbers, rebuses are defined as instances of entire *words* replaced by numbers. Examples include the words *too* and *to* replaced by the number *2*. To identify rebus patterns, all instances of the relevant numbers – they turned out to be *2* and *4* only – were extracted from the corpus and inspected in context. A search algorithm that extracted occurrences of *4* replacing *for*, *2* replacing *to* and *2* replacing *too* was created. This script tagged relevant text snippets for parts of speech, using the Perceptron tagger implementation in NLTK (NLTK Project 2013). The search algorithm took into account the lexical item preceding and following the *2* or *4* as well as the part of speech-tag associated with that item. For instance, if an item following an instance of *2* was (correctly) tagged as a preposition, the *2* is most likely not a representation of *to*. If the item was (correctly) tagged as verb, the *2* might very well be replacing a *to*. (Consider *Should have 2 to 4 years of experience* versus *Looking forward 2 seeing you.*) A word list was used to resolve ambiguous taggings; if, for instance, several tokens of *seeing*, as in the second example, were mis-tagged as prepositions, the word list

was used to instruct the algorithm to disregard the preposition-tag if it encountered the item *seeing*. All positives as well as excluded items were inspected in context.

The resulting search for *4* instead of *for* yielded 748 tokens. *2* replacing *to* occurred 393 times, and 155 tokens of *2* were replacing an instance of *too*.

Below, the most common contexts of this feature are listed, grouped by preceding word (word *4* ...) and following word (... *4* word). For instance, the combination *pic 4* ... occurred 226 times in the corpus.

Item preceding <i>4</i>	Tokens
<i>Pic</i>	226
<i>looking</i>	196
<i>Lookin</i>	35
<i>Pix</i>	23
<i>pics</i>	18

Table 12: Counts of items preceding *4* replacing *for* in the dating corpus.

Item following <i>4</i>	Tokens
<i>Pic</i>	209
<i>A</i>	120
<i>U</i>	27
<i>real</i>	21
<i>someone</i>	19

Table 13: Counts of items following *4* replacing *for* in the dating corpus.

The leet version 2 for *to* occurred 393 times. Again, the most common collocations are listed below.

Item preceding 2	Tokens
<i>looking</i>	33
<i>love</i>	25
<i>want</i>	25
<i>wants</i>	16
<i>has</i>	14
<i>ready</i>	10

Table 14: Counts of items preceding 2 replacing *to* in the dating corpus.

Item following 2	Tokens
<i>be</i>	35
<i>suck</i>	18
<i>meet</i>	16
<i>play</i>	12
<i>get</i>	12
<i>have</i>	11
<i>her</i>	11

Table 15: Counts of items following 2 replacing *to* in the dating corpus.

The substitution 2 for *too* was found 155 times.

Item preceding <i>?</i>	Tokens
<i>ub</i>	86
<i>b</i>	19
<i>be</i>	13
<i>that</i>	6

Table 16: Counts of items preceding *?* replacing *too* in the dating corpus.

Item following <i>?</i>	Tokens
<i>I</i>	28
<i>looking</i>	11
<i>much</i>	6
<i>hit</i>	6
<i>im</i>	5

Table 17: Counts of items following *?* replacing *too* in the dating corpus.

3.3.1.7. Feature 7: repeated punctuation

Instances of non-Standard punctuation were identified by compiling a regular expression based on the list of punctuation characters contained in the Python object *string.punctuation*, which contains all punctuation markers recognized by the software (Python Software Foundation 2017). This regular expression counted all instances of more than one consecutive item of punctuation as well as instances of non-Standard combinations of *?* and *!*, such as *!?!.*

Feature	Instances in the dating corpus
.	80,090
!?	12,093
-	3,164
,	2,223
*	1,754
?	1,709
+	1,366
'	541

Table 18: Counts of non-Standard punctuation in the dating corpus.

3.3.1.8. Feature 8: prosody

This category includes all typographic representations of non-linguistic emotional expressions. A dictionary of all non-Standard words found in the corpus was inspected to identify potential features in addition to the ones listed by Herring (2012:2). Tokens were collected using a regular expression on untokenized text.

Prosodic item	Tokens	Examples
<i>Ha</i>	1,010	<i>hah</i> <i>hahahahaha</i> ...
<i>Ya</i>	1,006	<i>Ya</i> <i>yah</i> ...
<i>So</i>	383	<i>Soo</i> <i>sooooooooooooo</i> ...
<i>Um</i>	153	<i>um</i> <i>ummmmmmmmmmmmmmmmmmmmm</i> ...
<i>Hm</i>	149	<i>Hm</i> <i>hmmmmmm</i> ...
<i>Mm</i>	146	<i>mmmmmmmm</i> <i>mmmm</i> ...
<i>Er</i>	84	<i>er</i>
<i>Hu</i>	73	<i>huh</i> <i>hu</i>
<i>He</i>	71	<i>heheheh</i> <i>hehehehe</i> <i>hehehe</i> <i>hehe</i>
<i>Hey</i>	61	<i>heyyy</i> <i>heyyyyyy</i> <i>heeyyy</i> <i>heeeeeeyyyyyy</i> ...

Table 19: Counts of prosodic items in the dating corpus.

3.3.1.9. Feature 9: replacing words with letters

This feature describes the use of single characters for a longer word based on phonetic similarity. It includes items such as *C* or *c* for *see* and *U* or *u* for *you*. A similar approach to the rebus finding algorithms was used, where the extraction algorithm was refined until all results matched the requirements. Fixed expressions that only occurred with the shortened version, such as *Guns n' Roses* or *Rock n' Roll*, were not counted as instances of replacing words with letters as the non-replaced form does not exist in the corpus.

Feature	Tokens	Short for
<i>u</i>	5,452	<i>you</i>
<i>n</i>	2,580	<i>and, in</i>
<i>U</i>	851	<i>you</i>
<i>b</i>	601	<i>be</i>
<i>r</i>	476	<i>are</i>
<i>N</i>	145	<i>and, in</i>
<i>B</i>	131	<i>be</i>
<i>R</i>	86	<i>are</i>
<i>c</i>	46	<i>see</i>
<i>x</i>	26	<i>ex (e.g. in ex-wife)</i>
<i>X</i>	14	<i>ex (e.g. in ex-wife)</i>
<i>C</i>	6	<i>see</i>

Table 20: Counts of single letter replacements in the dating corpus.

3.3.1.10. Feature 10: word level

To analyze differences in word usage, the corpus was tokenized using the NLTK `word_tokenize` function. After lower-casing all words, 80387 lexical types remained. The table below shows the ten most frequent words.

Word	Tokens
<i>i</i>	501,899
<i>and</i>	365,929
<i>a</i>	359,874
<i>to</i>	346,408
<i>you</i>	192,208
<i>the</i>	166,012
<i>for</i>	165,592
<i>in</i>	125,510
<i>looking</i>	111,968
<i>with</i>	110,418

Table 21: Word counts in the dating corpus.

3.3.2. Discovery

After all the linguistic features listed above have been extracted, it remains to investigate the writers’ linguistic performance with regard to these features. As outlined above, the first step is to group the ads by linguistic features, using a clustering algorithm. (Technical details of the process are given in the sections on Clustering algorithm, 3.3.2.1, Distance Metric, 3.3.2.2). This will help us to first identify what linguistic features are relevant in this dataset and what groups differ in their use of said features.

From a theoretical perspective, this approach means that the grouping of data will be based on linguistic criteria alone (Horvath 1985; Bamman, Eisenstein & Schnoebelen 2014). No language-external characteristics of the ads are taken into account at this point.

Based on this linguistic grouping of the ads, we can analyze to what extent patterns in this structure co-occur with social characteristics of the authors such as their gender. A similar clustering approach is taken in gender variation studies by Horvath's (1985) study of vowel variation in Sydney and more recently by Bamman et al. (2014) who work with Twitter data. The merits of this technique are acknowledged for instance by Russell (1987) and Milroy (1988).

The exploratory clustering presented here involves creating, inspecting and comparing multiple clusterings to see which clustering algorithm works best with the dataset (section 3.3.2.1.) and which distance metric is appropriate (section 3.3.2.2.) to compute the similarity of data points in the dataset. These issues will be discussed in the two sections below.

3.3.2.1. Clustering algorithm

The central idea of clustering is to group data points according to shared features, by computing their similarity to other points in the dataset. Thus, a clustering algorithm will take a number of data points – in this study, the dating ads – and divide them into clusters. The end result of this process will be called a clustering. Thus in the following, a *clustering* refers to a group of *clusters*, which in turn are groups of *data points*. Each data point represents a dating ad from our corpus. To achieve such a clustering of the dataset, various algorithms can be employed.

For this study, the nine clustering algorithms implemented in the Python machine learning module Scikit Learn (Pedregosa et al. 2011) were tested. (A complete list is given in the documentation (scikit-learn developers 2016)). Starting with a small test set, the size of the dataset was incrementally increased to test each algorithm's performance. Any algorithm that was unable to deal with a dataset of 40,000 ads and 300 features was

excluded. This applies to Spectral clustering, Mean Shift, Affinity Propagation, Birch Clustering, and DBScan.

The remaining clustering algorithms that were employed at some point in this study are outlined below; they are k-means, hierarchical clustering, and Gaussian mixture models.

The *k-means* algorithm, described in Crawley (2007:738) or Duda et al (2012:718–750), is one of the most widely used clustering algorithms. Supplied with a number of features n , and a number of clusters k , it locates each data point in n -dimensional space. For each point, it calculates the distance to its neighbors and then groups neighboring points into clusters. More specifically, the algorithm starts by creating k cluster centers (*the means*) and grouping each data point with the center closest to it. The algorithm then computes the error term of the result, and changes the position of the means, attempting to minimize this error term. Once no more improvement is seen, the resulting clustering is returned. K-means is comparatively efficient when dealing with large datasets (Duda, Hart & Stork 2012:561). However, the number of clusters needs to be specified in advance, so it is not as flexible as other clustering algorithms. K-means iterative approach – trying out constellations until no more improvement is seen – can also result in the algorithm ending up at a local maximum, rather than the actual best solution. This can be mitigated by running k-means several times with different starting constellations and pick the best solution.

Hierarchical clustering algorithms (Crawley 2007:742) are another group of distance-based algorithms (Duda, Hart & Stork 2012:550–55) . The agglomerative hierarchical clustering algorithm implemented in scikit-learn starts by treating every data point as its own cluster. The algorithm then successively merges neighboring data points into this cluster until it ends up with one big cluster that contains the whole

dataset. This results in a tree-like structure which the user can “slice” at any height to see a clustering with a desired numbers of clusters. Agglomerative clustering can either use the biggest distance, the average distance, or the so-called Ward distance between two data points when computing cluster membership. The difference between the two models described above and a *Gaussian mixture model* clusterer (Duda, Hart & Stork 2012:521) is that the Gaussian does not rely on distances between data points, but instead assumes that the data points come out of a group of different normal distributions. It iteratively computes the likelihood of the data under changing assumptions which distribution each data point is part of. That is to say, in the process, each data point is matched to one of the normal distributions; the algorithm then computes the probability of this specific data point originating out of that distribution. The data point is finally assumed to be a member of the distribution with the highest membership probability. This distribution then is this point’s cluster.

To be able on inspect and evaluate the resulting clusterings, a Python module called `clustertools` (available at <https://github.com/patrickschu/chapter2/blob/master/current/clustertools.py>) was created. It contains several classes and functions to help compute and visualize characteristics of clusterings as well as individual clusters. The goal was to produce the tools necessary to display the clustering results following the guidelines for a useful general cluster interface as specified in Grimmer & King (2011). Grimmer and King (2011:6) suggest that the main features of such a general cluster interface include

- 1) a breakdown of characteristics of the items in each cluster. For the present study, for example: how many male-written and female-written ads are in cluster X, in cluster Y, etc.?

- 2) the display of features that distinguish one cluster from the other clusters. For the present study, for example: what micro-level feature distinguishes cluster X from cluster Y, etc.?
- 3) an overview of how different clustering algorithms divide the data. For the present study, for example: how does the clustering produced by k-means differ from the clustering produced by a Gaussian mixed model?

These concepts are implemented in the `clustertools` module and allow us to output for each clustering the following:

- 1) Clustering statistics. How many clusters, that is groups of data points, does this clustering contain? How many items are contained in each cluster?
- 2) Intra-clustering statistics. For each cluster in a clustering: what proportion of the overall data does the cluster contain? How many items of each category are contained in the cluster?
- 3) Clustering quality. The most frequent clustering evaluation metrics are computed: The silhouette score (Rousseeuw 1987) measures cluster density by a ratio relating the intra-cluster distance to the distance to the neighboring cluster. The Jaccard index (Real & Vargas 1996) computes an accuracy ratio by dividing the number of identical items in two clusters by the total number of items. Implementations from Scikit Learn were used for all metrics described here (scikit-learn developers 2017b).
- 4) Inter-clustering comparison. If several clustering have been produced, metrics which indicate the similarity between clusterings are computed: the Jaccard index, the silhouette similarity (see above), and the rand index (adjusted and non-adjusted), which compares two clustering by considering the number of items that are grouped into the same cluster in both clusterings. Thus, each of

these computes a measure of the overlap of two clusterings. In our output, this helps us establish to what extent the different clustering algorithms concur on what is a sensible clustering of the data.

(A sample output is reproduced in appendix A.1.)

3.3.2.2. Distance metric

One last technical question needs to be considered: the choice of a distance metric. Distance-based clustering algorithms such as k-means need a way to measure similarity between data points in order to group data points close together into the same cluster. There are various distance metrics to choose from to compute this distance; the three most common ones, relevant here, are listed below. The *Euclidean distance* (Duda, Hart & Stork 2012:187) is the way one would intuitively measure the distance between two points: The length of a straight line between point A and point B. The *Manhattan* or city block distance (Duda, Hart & Stork 2012:188), on the other hand, measures distances the way one would when walking through a city: not as a straight line (“as the crow flies”) between two points but as a combination of rectangular straight lines around every block (“as the human walks”). The *cosine* similarity (Duda, Hart & Stork 2012:541) measures the cosine of an angle between the vectors of two points to establish their similarity. This metric, common in information extraction, has been found to be less sensitive to the distortions of high-dimensional data (Aggarwal, Hinneburg & Keim 2001) In this paper, the distance metrics as implemented in Python’s *SciPy* (Scipy community 2017) module were used because the toolkit offers a wide variety of metrics that integrate well with the scikit-learn tools.

3.3.2.3. Clustering micro-level features

In the following two sections, clusterings based on e-grammar micro features and based on overall word are described. The first sub-section below focuses on the micro e-grammar features, a description of the word-based clustering is given in 3.3.2.4

Just to re-iterate, the goal is to identify how the data is structured based on linguistic features. That is, does it fall into two groups, as a gender-based study would assume? Does it cluster as four groups, where the ad category or sexual orientation could be expected to be most relevant? Does it not display any kind of structure? Etc.

To prepare the ads for clustering by e-grammar features, each ad was converted into a vector containing frequencies (tokens per total words) of the 9 features. The example below shows the output for a 20-word ad, containing 1 alphabetism, no non-Standard capitalization, 2 items of leetspeak, etc.

[1/20, 0/20, 0/20, 0/20, 2/20, 0/20, 0/20, 0/20]

This approach resulted in very sparse matrices, that is to say most of the numbers were zeros – especially in generally rather uncommon features such as leetspeak. To address this issue and to also a preliminary gage of each feature’s relevance, a term weighting method called *tf-idf* (short for *Term frequency, inverse document frequency*) was computed for each ad. The *tf-idf* ratio gives an indication of each feature’s relative importance. More specifically, the “term frequency” – how often does feature X occur overall? – and the “inverse document frequency” – in how many documents does feature X occur? – are computed and then turned into a ratio (scikit-learn developers 2017c). The *tf-idf* ratio thus weights each feature according to its importance. Let’s assume, for instance, that feature X and feature Y both occur 500 times in a corpus of 500 ads. However, feature X occurs once in every ad (similar to a function word like *the*) while feature Y is concentrated in only ten ads (similar to a content word). Y will score higher

on the tf-idf ratio. The sociolinguistic equivalent would be a phonetic variant that is used frequently, but only by a specific group of the population. The tf-idf ratio thus allows us to see which features vary more strongly than others. It also controls to some extent for the wide divergence in frequency between the features – compare the 42 tokens of leetspeak with the thousands of instances of non-Standard punctuation. In the present study, the tf-idf implementation from Scikit-Learn (scikit-learn developers 2017a) was used with the standard settings except that a “smoothing” was de-activated; this procedure is needed in case a token does not occur in the corpus at all, which is not possible with the present approach (this is mainly relevant when running the model on unseen text). Setting the tf-idf’s “norm” parameter to “l2” resulted in the resulting tf-idf vectors being normalized in order to avoid giving long texts more weight, using the Euclidean distance norm.

Results for all 103,290 ads were fed into several clustering algorithms. The full dataset was clustered using hierarchical clustering, the Gaussian mixture, and the k-means algorithm. Each was run with settings to produce clusterings of 2, 3, 4, 5, and 6 clusters. These were then inspected using the cluster tools described above.

Based on our hypotheses, the two cluster and four cluster clusterings were given most attention: a 2-cluster solution would suggest that author gender or addressee gender were the social attributes most relevant to linguistic variation. A 4-cluster solution would point toward the categories (i.e. m4m, m4w, w4m and w4w) as important predictors of linguistic variation. In each case, inspection of the makeup of each cluster would give more insight into the dynamics. Using the standard k-means algorithm with Euclidean distance and ten initializations, the clustering with four clusters looked more promising. It scored higher on the silhouette metric (4-cluster clustering: mean silhouette coefficient = 0.42, 2-cluster clustering: mean silhouette coefficient = 0.33), suggesting

that it has better defined clusters. Looking at the composition of these clusters, the two clusters in the 2-cluster clustering do not pattern along binary gender lines: one cluster contains 88 % of male and 82 % of female authors, the other one 12 % and 18 % respectively. The 2-cluster clustering, that is, exhibits a weak clustering with gender groups distributed evenly over clusters. The 4-cluster clustering distributes the data points over the four clusters as follows: Cluster 0, 12 % of the total; Cluster 1, 53 % of the total; Cluster 2, 17 % of the total; Cluster 3, 18.0 % of the total. The data, that is, are not distributed evenly over clusters. Rather, more than half of the data ends up in Cluster 1. Looking at the distribution of categories over clusters, we see that it deviates from a purely random distribution, where the percentage from each category would mirror the percentage of the total data contained in the cluster. The m4m authors, for instance, are under-represented in the big Cluster 1 and over-represented in Cluster 2. The w4m writers, conversely, are under-represented in Cluster 2 and over-represented in Cluster 3. The m4w and w4w writers pattern pretty much the same.

Cluster	Percentage of dataset	Proportion m4w	Proportion m4m	Proportion w4m	Proportion w4w
0	12 %	12 %	13 %	12 %	11 %
1	53 %	54 %	47 %	57 %	57 %
2	17 %	14 %	26 %	11 %	14 %
3	18 %	19 %	14 %	21%	19 %

Table 22: Results of k-means clustering: cluster size and cluster membership by ad category.

Looking at the linguistic features correlated with cluster membership, the results suggest that the use of Capitals and to a lesser degree Single Letters is positively correlated with membership in the large cluster 1; Cluster 1 is noticeable for high use of abbreviations; the use of clippings is somewhat predictive of membership in Cluster 2; members of Cluster 3 are distinguished by relatively higher frequencies of non-Standard punctuation. The details need not concern us here, but the clustering allows us to state that

- 1) there is variation in the data that the clustering algorithm can pick up on; and
- 2) an approach that looks at four different groups appears more promising than investigating a binary gender split.

Thus, the detailed analysis of the micro-level features presented in the section 3.3.3.1. – after the description of clustering by word use below – will take this first finding as its starting point.

3.3.2.4. Clustering by word use

Below, the results of clustering ads based on word usage, rather than micro e-grammar features, are described. The procedure mirrors the approach to clustering micro-level features described above. Just to re-iterate, the goal is to identify how the data is structured based on linguistic features, in this case words: does it fall into two groups, as a gender-based study would assume? Does it fall into four groups, where the ad category, that is sexual orientation, could be expected to be most relevant? Does it not display any kind of structure? Etc. An initial analysis with clusterings based on individual words (bag of words model: all words over a certain frequency threshold) did not lead to satisfying results as essentially all ads ended up in one cluster. To enrich the model, a word2vec model (Mikolov & Dean 2013) was trained on the dataset. Word2vec translates each word into a vector of numbers, so-called *word embeddings*, based on other words it often co-occurs with.

For this study, a vector consists of 100 items was created for each word. Since word2vec is a neural network model, each items corresponds to a weighted node that can be turned on or off. During training, the neural network model learns the probability of a word based on the occurrence of other words a within a specific window; in the present study, the window size was set to 11 words. Suppose, for instance, that the word2vec algorithm is confronted with the following sentence, where the window is set to five lexical items (in brackets) and the word X is to be predicted:

This [is the] X [that I] love.

A successful word2vec model learns that words like *cat* and *dog* are more likely to be in position X than for instance *any* or *the*. Consequently, the vector for *cat* is going to be more similar to *dog* than to *the*; a word like *fire truck*, on the other hand, will be more similar to *cat* than *the* but not quite as similar as *dog*.

Here, the word2vec model was trained on the whole dataset (including the “Strictly Platonic” section to supply more training data), while excluding all words that occur less than 20 times. The order of the words in the 11-word window was taken into account, using so-called skip-gram training. (As opposed to a bag-of-words model, that only checks whether a word is present within the window or not). A k-means clustering was applied to the resulting vectors. This resulted in 54 groups of words, each forming a unit resembling a semantic group.

Each group was then manually assigned a label summarizing its contents. For instance, a group including terms such as *waist, tall, eyes, eyed, skinned, thinning, strawberry, tattoos, built, build, weigh, scruffy, cut* was named “physical characteristics”, one including *paso, wayne, palm, santa, colorado, richmond, houston, los, ohio, ca, louisiana*, was labeled “locations”. Four groups of words could not be assigned a consistent label. (The full document with all groups can be found in the appendix.) Starting with 54 groups and excluding the 4 non-labeled one, this leaves us with 49 features to extract. Essentially, this means that the number of words from each semantic cluster, rather than each word by itself, was extracted and used in the clustering.

Based on the experience with clustering e-grammar micro features, the clustering was achieved using the k-means algorithm with Manhattan distance and four clusters. The results are presented in table 23 below.

Cluster	Percentage of dataset	Proportion m4w	Proportion m4m	Proportion w4m	Proportion w4w
0	52.1 %	30 %	14 %	30 %	26 %
1	21.7 %	30 %	14 %	29 %	27 %
2	17.4 %	16 %	39 %	17 %	28 %
3	8.7 %	2 %	89 %	1%	8 %

Table 23: Clusters as percentage of dataset and breakdown by category.

We see that the groups are distributed unevenly over clusters in apparently non-random fashion. Thus, the analysis of the word level features presented in section 3.3.3.2. below will take this clustering as its starting point.

3.3.3. Analysis

3.3.3.1. Analysis micro-level e-grammar features

After clustering the data as described above, each feature needed to be analyzed individually. Parallel to the clustering described above, a binary (men and women authors) and a four-group comparison (the four categories m4m, m4w, w4m, w4w) were plotted and statistically analyzed. This gives us further insight into the differences between the approaches and will help evaluate the outcome of our cluster analysis, which suggested that a 4-way comparison rather than a 2-way comparison would be most fruitful. Figure 5 below shows the difference in the use of each feature by male and female authors. In this figure, the mean for male authors and female authors is plotted against the overall mean for each feature. This distance to the overall mean is measured in

standard deviations to make results comparable across features. Figure 7 suggests that there is no noticeable difference between those groups: both male and female authors stay close to the mean, except for clippings and emoticons, where we might spot trends towards more frequent usage among men and women authors respectively. This mirrors the result of our k-means clustering which did not discover any strong patterns along the male-female dimension.

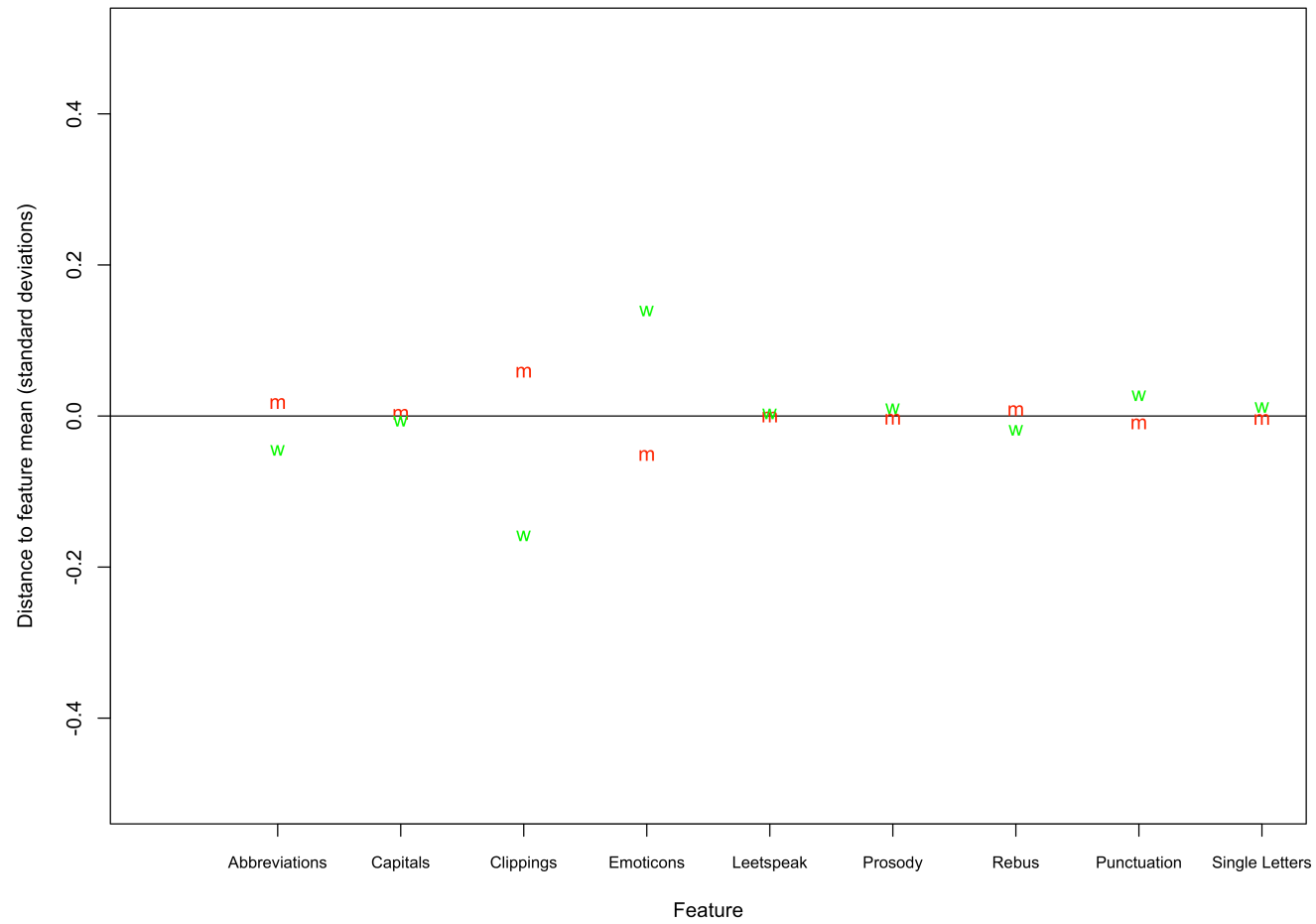


Figure 5: Frequency of features in the dating ad corpus, male versus female authors.

It remains to compare the results presented in Figure 5 to a plot based on the 4-cluster clustering. For this plot, we assume, based on our inspection of clusters in section 3.3.2.3., that these four groups correspond to the four categories in the dataset: w4w, w4m, m4w, m4m. The results are visualized in two plots (Figure 6, Figure 7) below. The first box plot shows the relative ranking of the four groups regarding feature frequency. The second one shows the same results in a more detailed manner and can be used to further illuminate insights from Figure 6.

The first boxplot (Figure 6) ranks the groups by feature frequency. It indicates, for instance, that m4m authors exhibit the highest per-word-frequency of abbreviations and capitalized words in the dataset. The m4w authors, on the other hand, were last in abbreviation use and second-to-last in use of capitalized words.

Two patterns can be identified here: One is that the m4m writers consistently occupy extreme positions, being the first-ranked group in six of nine and the last-ranked in two other categories. The w4w writers tend to be among more prolific users of e-grammar features as well. The w4m and m4w writers, on the other hand, consistently occupy low ranks – with two exceptions: prosody for m4w and emoticons for w4m. These are incidentally also the linguistic categories where m4m authors are not leading, but lagging behind. These patterns will be easier to interpret after also attending to Figure 7, which shows the results in more detail.

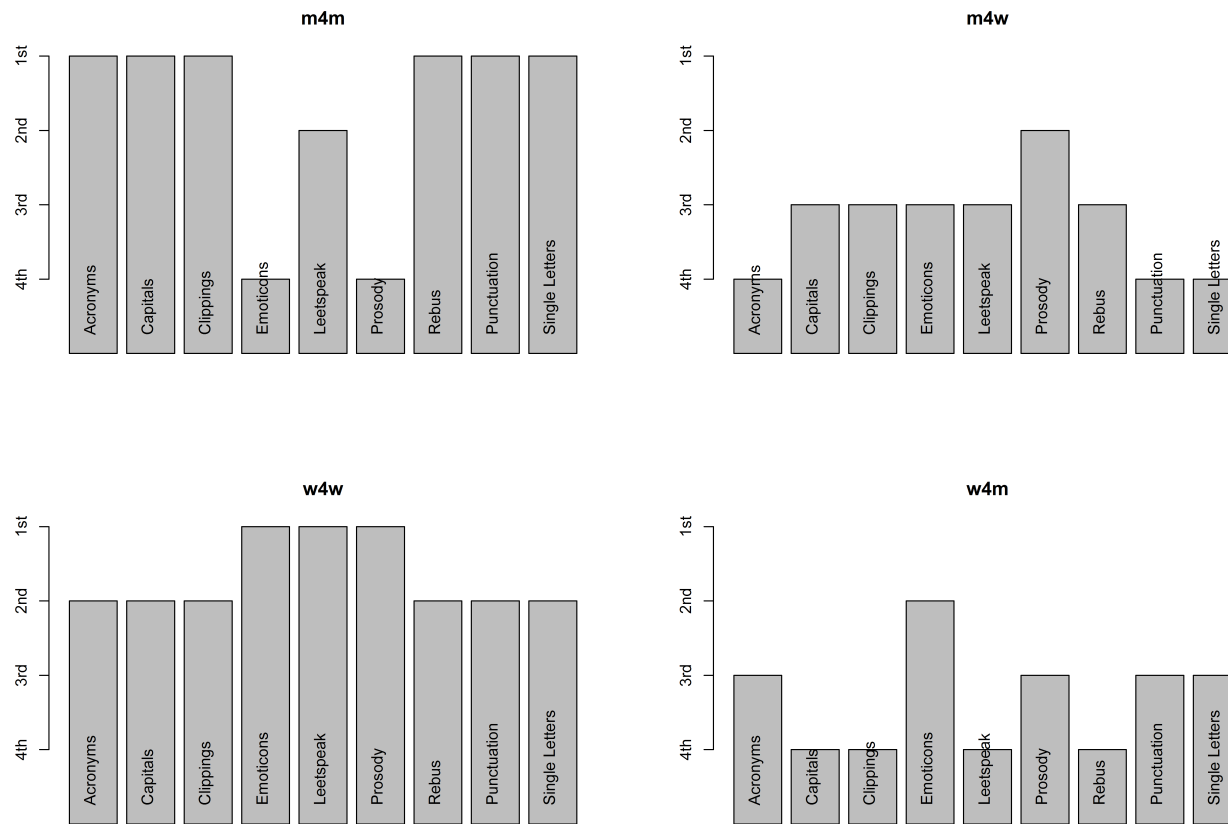


Figure 6: Ranked feature frequency by category.

When reading Figure 7, it is important to note that these are the same results presented in Figure 6, except that mean frequency (tokens per 1 million words), rather than rank, is plotted for each group. That is, it does not only show authors in which group on average use a feature the most or least, but also how much more or less often they use it compared to other groups. These differences are expressed as distance from the overall mean, measured in standard deviations. A data point above the $y = 0$ line (the zero line) indicates that members of this group use a certain feature more than the average author in the dataset. Data points below the line indicate that this group uses the feature less than the average author. The distance from the $y = 0$ line to the data point represents the strength of this effect. When it comes to abbreviations, for instance, the m4w writers use this feature a lot less than the overall mean. The same is true for the w4m writers, but the difference is not as pronounced, etc.

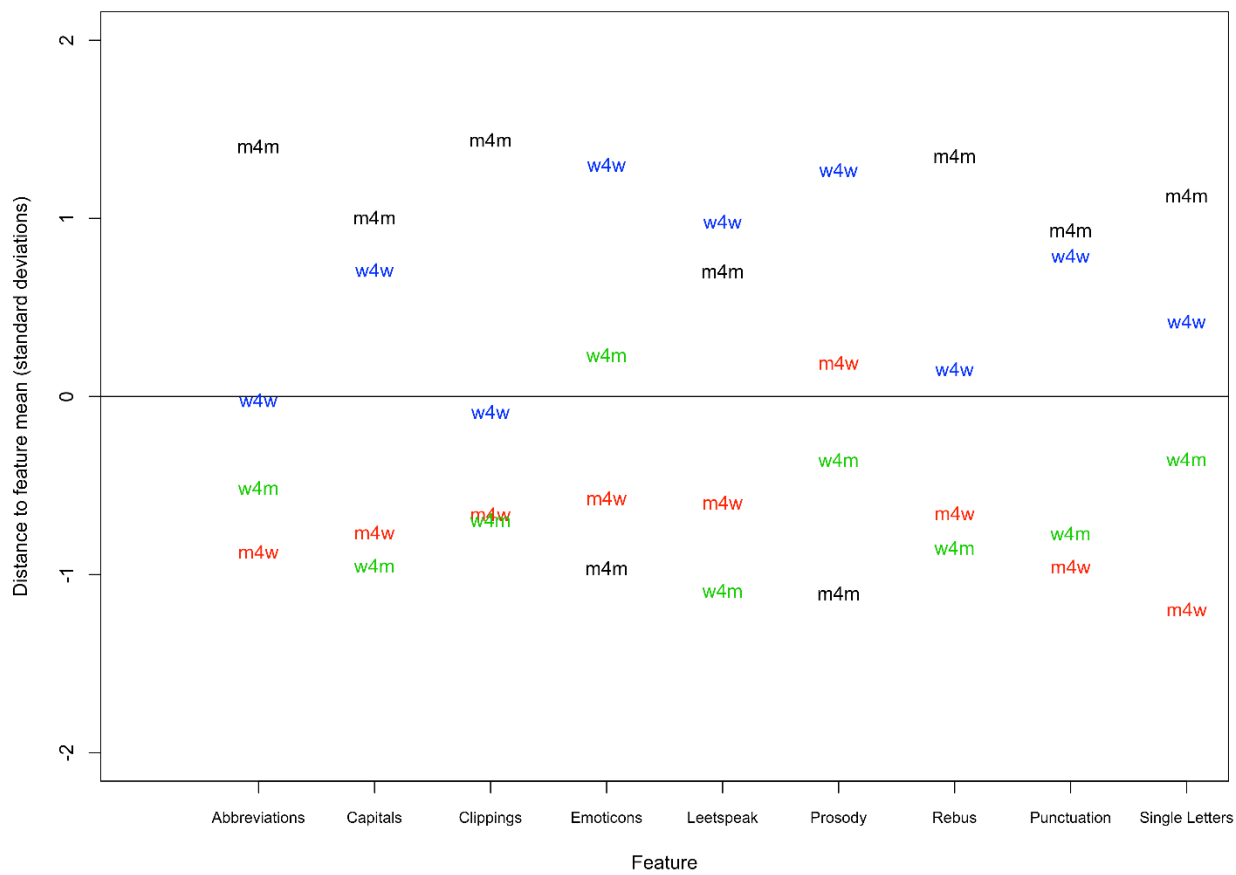


Figure 7: Relative feature frequency by category.

Figure 7 builds on Figure 6 and suggests that the data pattern in three distinct ways.

These three patterns can be characterized as follows:

Pattern 1: Data points are split into two separate groups, consisting of a m4m - w4w pair and a m4w - w4m pair. Examples: capitals, leetspeak, punctuation.

Pattern 2: Data points form a triad of w4m, m4w and w4w, with m4m as the outlier at the top. Examples: abbreviations, clippings

Pattern 3: Data points are evenly spread across all categories. Examples: emoticons, prosody, single letters.

It is also noteworthy that the m4m writers are above the mean or at the top in all features except prosody and emoticons. W4m and m4w writers, on the other hand, are below the mean for all features except prosody and emoticons. If we exclude these two features for now – a justification for this will be given below – we can subsume these three patterns as follows.

Pattern 1: m4m and w4w writers tend to use micro-level e-grammar features more frequently than w4m and m4w writers in this dataset.

Pattern 2: w4m and m4w writers tend to cluster together in their low use of micro-level e-grammar features.

Before we go on to discuss these patterns, it is important to consider why prosody and emoticons might pattern so differently in the plot above. To this end, (aspects of) the functionality of each feature are given in the table below. When “economy” is listed as a feature’s purpose, this indicates that the respective item can help a writer save characters or space, be it intentionally or coincidental. The other functions given are self-explanatory.

Feature	Example usage	Purpose
Abbreviation	<i>Looking for LTR</i>	Economy
Capitals	<i>NO WAY</i>	Add emphasis
Clippings	<i>Send a pic!</i>	Economy
Emoticons	<i>See you :)</i>	Add facial expression
Leetspeak	<i>S3nd a pic</i>	?
Prosody	<i>Hahaha</i>	Add non-linguistic expression
Rebus	<i>Looking 4 you</i>	Economy
Punctuation	<i>Let's see ..., Thanks!!!!</i>	Add implication, emphasis
Single letters	<i>U B 2</i>	Economy

Table 24: Functions of micro-level features.

This table is not intended to be a detailed semantic analysis of these features but it does illustrate a shared characteristic of prosodic features and emoticons: they add, rather than subtract, characters and they work on an extra-linguistic level by adding facial expression or non-linguistic sounds. We can thus expand on the statements presented as pattern 1 and pattern 2 above as follows:

Result 1: m4m and w4w authors tend to use micro-level e-grammar features more frequently than w4m and m4w authors in this dataset. Exceptions are features representing non-linguistic communication, namely emoticons and prosody.

Result 2: w4m and m4w writers tend to cluster together in their use of micro-level e-grammar features in this dataset. Their use frequency of features is low,

exceptions are items representing non-linguistic communication, namely emoticons and prosody.

These results will be discussed in more detail in chapter 5, after all features have been subjected to a matched guise perception test. Further details, such as the status of repeated punctuation, which does not quite fit in this schema, will be addressed at this point as well.

3.3.3.2. Analysis word use

The word2vec-based clustering presented in the Discovery section (3.3.2.) is reproduced here for convenience:

Cluster	Percentage of dataset	Proportion m4w	Proportion m4m	Proportion w4m	Proportion w4w
0	52.1 %	30 %	14 %	30 %	26 %
1	21.7 %	30 %	14 %	29 %	27 %
2	17.4 %	16 %	39 %	17 %	28 %
3	8.7 %	2 %	89 %	1%	8 %

Table 25: Clusters as percentage of dataset and breakdown by category.

It is noteworthy that the clusters are unevenly sized, ranging from cluster 0, containing 52 % of all ads, to cluster 3 with only 9 % of the data points. Looking at the representation of the four categories m4w, m4m, w4m, and w4w across the clusters, we find that:

- 1) Cluster 0 contains mainly w4m and m4w data; the m4m category is under-represented.

- 2) Cluster 1's distribution mirrors cluster 0.
- 3) Cluster 2 consists of m4m and w4w ads mostly; these categories account for almost 70 Percent of the data points in cluster 2.
- 4) Cluster 3 consists almost exclusively of m4m ads.

The distribution of categories over clusters is shown in table 26 below. The table illustrates patterns in the distribution of categories over clusters. It suggests that comparing m4m writers to the rest of the authors is the most relevant distinction in word use.

Again, in a random, non-stratified clustering, we would expect all data points to be distributed evenly over clusters. Thus, if a cluster contains 9 % of the total dataset, we would expect the proportion of w4w, m4m, w4m, m4w ads contained in this cluster also to approach 9 percent. In the table below, the proportion of each group is contrasted with the expected value; instances where the actual value substantially undershoots the expected value are in bold; an asterisk indicates that they are higher than would be expected for a random distribution.

Category	Cluster 0	Cluster 1	Cluster 2	Cluster 3
m4w	* 62 % (52)	* 26 % (22)	11 % (17)	1 % (9)
m4m	29 % (52)	12 % (22)	27 % (17) *	* 31 % (9)
w4m	* 62 % (52)	* 26 % (22)	12 % (17)	0 % (9)
w4w	55 % (52)	23 % (22)	19 % (17)	3 % (9)

Table 26: Results of k-means clustering, distribution of categories over clusters in percent.

The special role of the m4m writers is illustrated by an analysis of distribution by ad category in the plot below (Figure 8).

The plot illustrates the three main findings: first, the m4m ads are an outlier. They are consistently found at much higher or lower quantities in each cluster than would be expected. The categories w4m and m4w, on the other hand, are very close to overlapping in their distributions over clusters. The w4w ads fall somewhat in the middle but align more closely with the m4w and w4m ads. In terms of gender and sexual orientation, we can state that the analysis shows quite a few ads written by gay men to be quite different from the rest of the rest of the dataset. Heterosexual writers, be they male or female, tend to share a lot of lexical characteristics. Most gay women are slightly different from the heterosexual couples, but tend to differ in certain aspects as is shown in the analysis of distinctive features below.

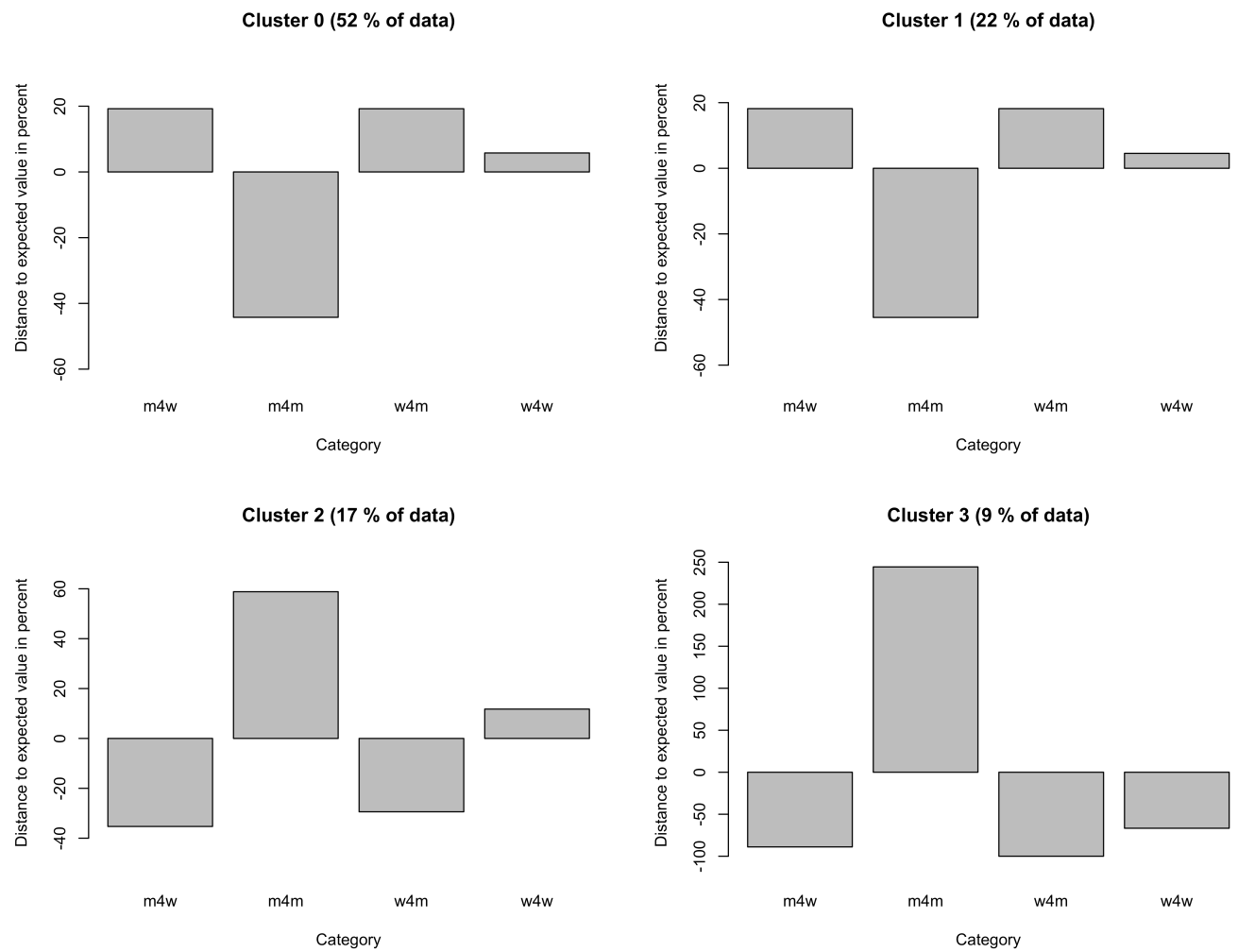


Figure 8: Cluster membership by category, distance to value expected in random distribution in percent.

The visualization in Figure 8 does not show, however, what linguistic features drive differentiation. That is, what are the features where each cluster differs the most from the three others? The table below illustrates which of the word2vec semantic groups are crucial in distinguishing each cluster from the other. The magnitude of the difference is expressed in z-scores; these are computed by subtracting means of all centroids from the current centroid and dividing by the standard deviation of all centroids for this feature. Thus, just like in Figure 7, they represent the number of standard deviations the data point is distant from the mean.

The table is meant to be read from left to right, similar to a confusion matrix. Starting with cluster 0, we see that when compared to cluster 1, it scores a + 3.04 in the semantic group “Physical Characteristics”, and a negative 3.77 when compared to cluster 2 as far as the score in the “Age & Ethnicity” category is concerned, etc. A negative value indicates that the cluster on the left side scores lower on this feature than the comparison cluster from the top row, a positive value indicates that members of this cluster tend to use words from a semantic group more often. For the example cited, this means that texts in cluster 0 tend to contain more words from the “Physical Characteristics” group, but fewer items from “Age & Ethnicity” when compared to cluster 1.

X	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Cluster 0	X	Physical characteristics: + 3.04 Jobs & Education: - 2.8 Drugs + 1.13	Age & Ethnicity - 3.77 Physical Characteristics: + 3.55 Leisure Activities: - 2.11	Sex Terms & Body Parts: - 5.55 Physical Characteristics: + 2.43 Positive Characteristics & Adjectives: -1.36 Drugs: + 1.3 Sex Terms & Body Parts: - 5.36 Jobs & Education: + 3.95 Positive Characteristics & Adjectives: - 1.21
Cluster 1		X	Age & Ethnicity: -3.48 Jobs & Education: 3.28 Leisure Activities: - 2.41 Positive Characteristics & Adjectives: - 1.19	

Table 27: Semantic groups indicative of each cluster.

Cluster 2	X	Sex Terms & Body Parts: - 5.21 Age & Ethnicity: + 3.5 Physical Characteristics: - 1.11 Leisure Activities + 1.5
Cluster 3	X	

Table 27: continued.

We see that the m4m-heavy cluster 3 is mainly distinguished from other clusters by high scores in the “Sex Terms & Body Parts” semantic field. The “mainstream” cluster 0 is differentiated from all other clusters by the less frequent use of words from the semantic field “Physical Characteristics”, a semantic field including terms like *bodybuilder* or *tattoo*. Cluster 1 is distinguished from the majority cluster 0 by more reference to “Jobs and Education” and less discussion of “Physical Characteristics” or “Drugs”. Cluster 2 similarly differs from the mainstream cluster 0 in regard to “Physical Characteristics” but tends to have higher scores on “Leisure Activities” as well as “Age & Ethnicity”. The “Age & Ethnicity” and “Leisure Activities” fields are also important differentiators between clusters 1 and 2, with cluster 2 scoring higher in both of them. Again, cluster 1 is relatively higher on “Jobs & Education” than cluster 2.

This suggests a genre difference between the m4m writers and the rest of the dataset: ads written by m4m authors tend to be more focused on explicitly sexual content, and less on career and educational background than the rest of the dataset. This lexical analysis might be an indicator that m4m dating ads constitute their own sub-genre within the dating ads context. Just as in the use of micro e-grammar features, the m4m writers are outliers in their linguistic behavior. Taking this one step further, we could understand the patterns in the production data to indicate that existence of three sub-genres of dating ads: m4m ads, w4w ads and w4m/m4w ads, what one might call the “heterosexual ads” genre. The implication of this finding will be discussed in chapter 5, after the features have been subjected to a matched guise perception test in chapter 4.

3.4. RESULTS

Based on the production study, we can make the following statements regarding the hypotheses formulated at the beginning of the chapter:

- H 1** There is variation by binary author gender (male/female) in the data.
- Result** Author gender is a weak predictor in this dataset. An explanation in terms of binary author gender only obscures possibly meaningful patterns.
- H 2** There is variation by binary addressee gender (male/female) in the data.
- Result** Similarly, addressee gender is a weak predictor in this dataset. None of the features pattern strongly by addressee gender.
- H 3** There is variation by the author's and addressee's sexual orientation (binary: heterosexual/homosexual) in the data.
- Result** H3 is true for several features. Several features where m4m and w4w authors lead in feature use can be interpreted this way.
- H 4** There is variation by an interaction of the above in the data.
- Result** Hypothesis 4 is supported by the results of the production study in non-straightforward ways. Feature frequencies vary quite strongly by author – addressee dyads: for instance, m4m authors differ linguistically from m4w authors. This suggests a linguistic accommodation effect or variation by sexual orientation. One general pattern that emerges out of the data is that men writing to other men tend to use e-grammar features other than emoticons and prosodic items more frequently than the other dyads. When writing for women, on the other hand, male authors are below the overall mean in the use of the very same features. Similarly, women writing for other women tend to use e-grammar features more frequently than women writing for men do. This applies for to abbreviations, capitalized words, and clippings.

The two features emoticons and prosody exhibit a strikingly different pattern: here, male authors writing for a male audience are on average the least frequent users, with women writing for other women being the most prolific users and w4m and m4w authors converging around the mean. Especially for these two items, a linguistic accommodation effect seems possible. This approach is more satisfying from a linguistic standpoint than arguing for a “lesbian register” or “gay slang”. It will allow us to investigate general population patterns by framing the problem in terms of speaker and audience – concepts that are well-described in sociolinguistic theory.

H5 The variation reflects American gender ideologies.

Result H5 will be addressed in chapter 5.

Chapter 4: Perception

This chapter presents the second part of the investigation, a linguistic perception study. It is based on the results of the production study presented in chapter 2 in that it presents the same linguistic features to 891 participants of a matched guise language evaluation survey.

The chapter will thus focus on the side of linguistic indexicality not reflected in the results of chapter 3: the *perceptual* value of e-grammar features. If, as defined in the criteria for H1 and H2 in section X, any features exhibit gender-linked effects in both production and perception data, they merit further investigation into their social – and potentially gendered – meaning. Correlation between a gender category and a linguistic feature alone is not telling us much. Rather, we also need to find out whether said feature has a gendered meaning to the audience who encounters it (Agha 2007; Campbell-Kibler 2010). Consequently, the production and perception study differ in response variable (use versus perception of linguistic items) and methodology (corpus study versus evaluation task), but are addressing the same issue – linguistic indexicality – from different angles. Consistent with the methodological premises outlined in chapter 1, the studies attempt to do so by analyzing the same features in the same context to make their results comparable.

The perception study, in addition to acting as a foil for the results of the production study, is also designed to reflect on gender ideologies held by study participants. This data will mainly be gathered from responses to the open-ended questions about stereotypically gendered attributes (“Men are ...”, “Women are ...”) and by testing for correlations between perceived author gender and other perceived author

attributes, such as friendliness. This data on gender ideologies will be crucial to the discussion of social meaning to follow in chapter 5.

4.1. THE RESEARCH QUESTION

This study will build on insights from the production study presented in chapter 2: the features identified as potentially relevant in the production study will be tested for their perceptual relevance here. This will help ascertain whether features shown to correlate with gender or specific social groups are actually perceived as meaningful by native speakers of American English. In other words: are the findings from the production study random variation, artifacts of the dataset such as genre effects, or are they something humans pick up on and attach meaning to?

In this study, participants are exposed to the various linguistic features discussed in chapter 2, such as emoticons or repeated punctuation. An online questionnaire is used to assess participant evaluation of each feature.

This perception study will attempt to do the following:

- 1) Establish e-grammar features relevant to gender perception.
- 2) Find out what other attributes participants associate with gendered features. This will give insight into the broader gender ideologies participants hold.
- 3) Relate features from 1) and 2) and the interactions between them to ideologies of femininity and masculinity.

We will thus be able to test the following hypotheses:

- H 1 Certain writing styles and linguistic features are perceived as masculine or feminine.
- H 2 Gendered features will have other social meanings (e.g. traits such as assertiveness) attached to them.

H 3 Gender perception will vary by the study participants' own gender, age, and other social characteristics.

H 4 Perceived author gender will interact with perceived author sexual orientation.

The approach to addressing these hypotheses is described below: an outline of previous literature and reference materials consulted is given in section 4.1.1.; in section 4.2., the design of the text stimuli and the questionnaire is described; section 4.3. presents the results of the perception study.

4.1.1. Previous literature

The approach taken here applies the *matched guise* technique, commonly employed in sociophonetic studies (e.g. Smyth, Jacobs & Rogers 2003; Campbell-Kibler 2008) to writing. Matched guise studies expose participants to a recording and ask them to record their impression of the speaker based on the recording they listened to. Typically, recordings of the same speaker performing a text in different accents or languages (different “guises”) alongside some distraction scenarios are played to study participants. After each recording, participants rate the speaker on qualities such as friendliness or education level. Afterwards, differences in ratings are analyzed: for instance, do participants consistently rate a speaker higher on friendliness when the speaker adopts a Southern accent?

Matched guise tests were initially used to explore participants' perception of and attitude towards different languages and dialects. Lambert et al. (1960), for instance, played recordings of French-English bilinguals speaking in English and French guises to 64 participants. Lambert et al. (1960) found that English guises were rated more positively on attributes such as “good looks” or “sense of humor”. A positive evaluation of

Received Pronunciation in contrast to other varieties of English was documented in similar fashion by Ball (1983), Huygens & Vaughan (1983), and Stewart (1985).

Gender-related linguistic matched guise studies include Uldall (1960), Edelsky (1979), Smyth, Jacobs & Rogers (2003), Levon (2007), and Campbell-Kibler (2008; 2009; 2011). Two recent matched guise studies (Queen & Boland 2015; 2016) adapted the technique to written language by using text stimuli. An overview of this research paradigm is given below.

Edelsky (1979) studied the social perception of word-final rising intonation. 30 participants were asked to rate speakers on scales such as “confident – not confident” or “cool – warm”. Edelsky found that speakers using word-final rising intonation were more likely to be associated with more stereotypically feminine attributes such as “not confident”. Uldall (1960) had 30 participants rate recordings of male and female speakers on attributes such as “bored – interested” and “polite – rude”, with very similar results to Edelsky’s (1979). Aronovitch (1976) played 57 audio samples to 50 participants and found that female speakers were perceived as less confident and more extroverted.

Several studies also tested which variants are perceived as “gay” by participants, that is for the speaker’s perceived sexual orientation. Smyth, Jacobs & Rogers (2003), for instance, had participants label 25 male voices as more or less “gay sounding”. Smyth, Jacobs & Rogers then studied how text genre interacted with the sexual orientation assessment, finding that straight speakers were more likely to be heard as gay when reading a scientific text than when doing a dramatic reading. They hypothesize that formal speech is more “gay-sounding” (Smyth, Jacobs & Rogers 2003:337). A second study reported in Smyth, Jacobs & Rogers (2003) suggests that gay ratings are correlated with “masculine – feminine” scores: a speaker that is perceived as gay tends to also be perceived as more feminine than the average speaker. In a study investigating perceived

sexual orientation in men, Campbell-Kibler (2011) manipulates pitch, /s/-realization and realization of word-final *-ing* in recordings played to study participants. 175 participants were asked for the speaker's perceived sexual orientation as well as perceived education level. The study indicates that /s/-fronting is perceived to be less masculine and more gay, while (ing)-variation correlates with perceived education: a speaker using the reduced form [ɪŋ] tended to be rated lower on education level than the same speaker using the Standard [ɪŋ] pronunciation. (This finding is replicated in Campbell-Kibler (2009)).

Levon's (2007) study similarly tests whether pitch range and sibilant duration are good predictors of the perceived sexual orientation of a speaker. He presents participants with digitally manipulated recordings of the one speaker. Levon does not find a correlation between variants and perceived sexual orientation. Levon sees this as evidence for the complex nature of indexicality: "[T]his result, or more appropriately the lack thereof, serves to reinforce the notion that indexicality is not a straightforward process by which particular linguistic variables (or clusters of variables) are linked with social positions" (Levon 2007:549). A similar point is made by Campbell-Kibler (2008) in a study discussing the various social meanings listeners attach to variants of word-final *-ing*. Those meanings include friendliness, education, and whether the speaker is "annoying", "trying (too hard)", or "compassionate". Campbell-Kibler (2008) shows how these attributes all intersect in their indexical meaning.

While not concerned with gendered variation, Walker et al.'s (2014) study of the perception of variants of Spanish /s/ across listener groups relates to the present study in that it analyzes the impact of participant demographics and stylistic context on social meaning. Walker et al. (2014) played recordings of Mexican and Puerto Rico-Spanish speakers to 167 participants, after manipulating recordings to include /s/-presence and

absence. They found that while all listeners share an indexical field for (s), slightly different social meanings are activated based on the dialect of the speaker they are evaluating: the impact of the non-prestige variant on perceived speaker status is much stronger for Mexican speakers, in whose dialect the prestige variant is the default, than for Puerto Rico Spanish speakers, where the non-prestige variant is usually used. Walker et al (2014:169) conclude that this suggests that “listeners integrate their own local ideologies with their understanding of regional differences when socially evaluating language variation.”

Few matched guise studies have been conducted with written data as stimuli. Two papers by Queen & Boland (2015; 2016) are the exceptions within the linguistics literature. Queen & Boland (2015) presented 30 participants with text stimuli, supposedly emails of a writer replying to a “house mate wanted” ad. These texts were manipulated to variably include a number of features usually perceived as typographic or grammatical errors. These errors fell into one of three categories: first, what Queen & Boland (2015) call keyboarding errors (commonly referred to as “typos”, such as <abuot> for <about>); second, so-called “grammos”, homophone errors that are only relevant in writing (for instance selecting the wrong item from the group *to/two/too*); and hypercorrections: errors that are not unique to written language and are generally not stigmatized, such as *I/me* variation in subject position (*Marc and I* versus *Marc and me*).

Participants were presented with three stimuli, containing typos, grammos, and hypercorrections respectively. Participants then rated the writer on 12 questions that were designed to measure the writer on a social (“This writer is similar to me”) as well as an academic scale (“This writer is intelligent”). Queen & Boland find that while typos and grammos negatively affect the “academic” score, the “social” score is significantly impacted by the grammos – such as *to* instead of *two* – only. Queen & Boland also find an

interaction between participants' use of online media and perception ratings: participants who communicated more online (for instance on Facebook) tended to be less bothered by typos and grammos. To identify the most salient among the typos, grammos and hypercorrections, Queen & Boland (2015) had participants edit a text that contained all of these errors. More than 60 percent of participants fixed typos, while the rate for grammos and hypos was below fifty percent, leading Queen & Boland to label typos the most salient type of error in their text. Queen & Boland (2015:283) argue that this "suggests that written errors, when they are salient, contribute to the social meaning of text".

In a second paper presenting results from this experiment (Boland & Queen 2016), Boland & Queen study how a participant's personality traits (based on a "Big Five" questionnaire participants had to fill out, assessing characteristics such as extraversion and agreeableness) influence participants' ratings: what traits make participants more or less sensitive to grammos and typos? Results of the personality questionnaire emerge as a stronger predictor of ratings than for example a participant's sensitivity to errors in the editing task or their self-reported regard for good grammar. For instance, participants who scored high on "agreeableness" were more likely to give the text author a positive rating even if the text contained errors. This, Boland & Queen argue (Boland & Queen 2016:12), shows that linguists need to pay attention to the characteristics of the reader or listener to understand the social evaluation of language.

4.2. METHOD

The perception experiment presented here differs from most of the matched guise studies cited above in that it works with written stimuli. It does, however, follow the

principles of matched guise (Giles & Billings 2004; Drager 2013) and questionnaire studies (Schleef 2013) in sociolinguistics in several ways.

First, participants are exposed to a stimulus in the form of a personal ad and asked questions that are designed to assess their perception of the author. Those questions about perceived author attributes include binary multiple-choice questions (e.g. *Is the author female or male?*) as well as open-ended questions (e.g. *Men are ...*) and Likert-scale questions that ask participants to rate the author on a 5-item scale (e.g. from *very educated* to *very uneducated*). Likert-type questions offer the participant a scale on which to rate the stimulus author and allow for more nuanced feedback (Maurer & Pierce 1998).

Different participants are exposed to different variants of the text (from here on: stimuli), each differing only in the use of one linguistic feature. After questionnaires are completed, the results are inspected for correlations between responses and stimuli, i.e. whether the average response to a question varies significantly between stimuli.

Overall, seven stimuli were created. They were compared to a version of the text that did not contain any of the features of interest (from here on: the control stimulus). For example, emoticons were added to the control stimulus to create the “emoticon stimulus”. Only features addressed in the production part of the study (see analysis of the dating ads corpus in chapter 2) were included in stimuli. The features rebus and leetspeak were not included since they were very infrequent in the dating ad corpus and time and money for this study were limited.

The main focus of this study is on what gender identity participants assign to the author and the addressee. However, other questions are included in the questionnaire to investigate broader themes of gender ideologies and to distract participants from the research objective. The following section describes the design of this questionnaire.

4.2.1 Creating the questionnaire

The survey was designed and conducted through the online survey service Qualtrics. Participants (200 for the control stimulus, a target of 100 for each of the other stimuli) were recruited on the website Amazon Mechanical Turk. A sample size of 100 as sufficiently statistically powerful was established using the power formula $(z\text{-score})^2 * \text{standard deviation} * (1 - \text{standard deviation}) / (\text{margin of error})^2$ (Smith 2013). The z-score was set to 1.96 to represent a desired confidence level of 95%, the default value of 0.5 was entered for the standard deviation and a margin of error of $+ .75\% / - .75\%$ was deemed acceptable. On the Mechanical Turk platform, the assignment was advertised as “Who wrote this? 3 minute survey”. Participants were shown, in this order,

- 1) a cover letter, giving some background on study goals and IRB approval
- 2) instructions:

The following text is intended to be posted on a social networking web site. Please read it and answer the questions on the next page. (Identifying information such as email addresses has been removed.)
Thank you!

- 3) the stimulus (see sections 4.2.2. and 4.2.3.)
- 4) one page of multiple choice and Likert-scale questions about the stimulus (see table 25 below)
- 5) one page asking for participant demographics such as age, gender, ethnicity, sexual orientation, current place of residence, and hometown.

The following questions covering the perceived attributes of the stimulus’ author were included:

Response Type	#	Question	Options	Variable
multiple choice: binary	1	The author is ...	<input type="checkbox"/> male <input type="checkbox"/> female	perceived author gender
	2	The author is ...	<input type="checkbox"/> homosexual <input type="checkbox"/> heterosexual	perceived author sexual orientation
	3	The author is writing for ...	<input type="checkbox"/> a man <input type="checkbox"/> a woman	perceived addressee gender
multiple choice: 4 items	4	I'd guess the author is ...	<input type="checkbox"/> Asian <input type="checkbox"/> Black/Afr.-Am. <input type="checkbox"/> White <input type="checkbox"/> Hisp./Latino	perceived author ethnicity

Table 28: Questions assessing author perception in the matched guise study: type of question, response options, and perceptual variable tested for.

Likert scale: 5 items	5	The author seems ...	educated	perceived author education
	6		assertive	perceived author assertiveness
<input type="checkbox"/> Very __ <input type="checkbox"/> Somewhat __				
	7		sensitive	perceived author sensitivity
<input type="checkbox"/> _				
<input type="checkbox"/> Not Very __,				
	8		friendly	perceived author friendliness
<input type="checkbox"/> Very un__.				
	9		attractive	perceived author attractiveness
	10	Would you reply to this ad?	likely	ad effectiveness
Open-ended	11	Men are...	NA	gender stereotypes
		Women are ...		
	13	Do you have any further comments? Please share!	NA	additional comments

Table 28: continued.

(A reproduction of the complete questionnaire can be found in the appendix).

These questions were designed to address all aspects of gender discussed previously: author gender, author sexual orientation, and addressee gender (questions 1 – 3). Question 4 was used to check for interactions with ethnicity. The ethnicity categories follow the guidelines of the National Center for Education Statistics (NCES 2017).

The Likert-scale questions (5 – 10) covered personal attributes often discussed as gendered such as assertiveness, or sensitivity (cf. Eckert & McConnell-Ginet 2013) as well as characteristics discussed in previous matched guise studies with an interest in gender (friendliness, education, e.g. in Campbell-Kibler (2008)). These not directly gender-related questions were intended to distract participants from the study’s goal as well as to see whether any of them interact with gender perception. They were implemented as 5-item Likert-type rating scales rather than binary items. This allows for more nuanced measurement, as participants can rate the along a range rather than pick one of two values. For further analysis, Likert scores can then be converted into numbers, allowing for comparison between stimuli by computing averages or running statistical significance tests.

The open-ended questions were intended to give insight into gender stereotypes and provide context to the ratings. They also offer participants the opportunity to alert the researcher to problems with the study design or share additional thoughts they have.

Questions were shown in random order. Only the “Men are ...”, “Women are ...”, and “Comments” fields were fixed as the final questions (in this order). Participants could skip questions without penalty. Participants could only participate in one round of the survey: after completing the questionnaire, they were excluded from further runs via Mechanical Turk’s “Qualifications” system. That is, anyone who had rated the emoticon stimulus was excluded from the survey on the clipping stimulus, etc. Participants were rewarded with 10 cents, later 15 cents, and then 20 cents for completing the survey. The

increase in reward proved necessary when later rounds of surveys were not completed in a timely fashion until more compensation was offered. Participants from outside the United States were excluded by setting the criteria on Amazon Mechanical Turk accordingly. This approach allowed for rather fast and comparatively cheap collection of survey data.

This questionnaire thus was designed to allow us to address

- 1) which features are perceived as gendered in some way (based on questions 1 - 3). This addresses H 1 presented above.
- 2) what other personal characteristics, such as friendliness, are perceptually associated with these features (based on questions 5 – 12). This addresses H 2 presented above.
- 3) how gendered features and perceived personal attributes interact. This addresses H 3.

4.2.2. Creating the control stimulus

Before running the survey, a feature-neutral stimulus, the “control stimulus”, and several stimuli with added features, the “treatment stimuli”, had to be designed. The control stimulus is described below; the creation of treatment stimuli is the subject of section 4.2.3.

The control stimulus serves as a neutral baseline against which results for other stimuli are measured. To design this control stimulus, a text following the schema for dating ads established by Coupland (1996) was created by the author. Form and content were modeled on frequent patterns in the dating ads corpus. The text included no gendered pronouns or indicators of the author’s or addressee’s gender; only activities that could be considered gender neutral were included. The text did not include any

reference to sexual activities or the type of relationship sought as these seemed to be perceived very male-gendered in test runs. Care was taken to create the text in a way that allowed for introduction of relevant features. For instance, the words *information* and *especially* were included so that they could be clipped into shorter versions (*info, esp*) to create the clipping stimulus. Several drafts of the control stimulus were subjected to test runs with 30 participants until a version without too strong of an initial gender bias was identified. The final version of the control stimulus was rated by 200 participants on Mechanical Turk. The version used in the study is reproduced below.

Hi there,

I'm looking for fun people to hang out with from time to time, especially on the weekends.

Going to the beach, going to the movies or having a drink. Or just explore the city. If you're interested, contact me at @gmail.com for more information.

Looking forward to hearing from you!

The questionnaire for the control stimulus did not include the questions about author attractiveness and the “Men are ...”, “Women are ...” questions. Those were only added after the control stimulus run had been completed as new research questions were developed.

4.2.3. Creating treatment stimuli

Treatment stimuli were created by adding tokens from one of the relevant categories – capitals, clippings, emoticons, prosody, punctuation, single letters – to the control stimulus. Results from these stimuli could then be compared to the control stimulus. Two tokens of the respective feature were added, e.g. two emoticons or two instances of non-Standard punctuation. The six stimuli created and rated by 100 participants on Mechanical Turk are summarized in table 29 below. The emoticon stimulus was tested on 200 participants as the first version accidentally did not include the question about author attractiveness and the “Men are...”, “Women are...” sections.

Stimulus	Feature	Tokens changed or added
1	Capitals	<i>IF YOU'RE INTERESTED, CONTACT ME AT</i>
2	Clippings	<i>esp on the weekends [...] more info</i>
3	Emoticons	<i>Hi there :) [...] explore the city ;)</i>
4	Prosody	<i>explore the city, haha [...] Soooo if you're</i>
5	Punctuation	<i>Explore the city... [...] for more information ...</i>
6	Single letter	<i>If <u>ur</u> interested [...] to hearing from <u>u</u></i>

Table 29: Linguistic tokens added to or changed in the control stimulus to create the six treatment stimuli.

4.2.4. Method evaluation

The approach presented here tries to strike a balance between covering as many features as possible while still working with substantially-sized samples: 200 participants rated the control stimulus and 100 participants rated each of the 6 treatment stimuli. (The emoticon stimulus had to be administered twice because the first questionnaire posted was incomplete, thus there are 197 responses for this stimulus). This resulted in a total dataset of 891 completed questionnaires.

The survey methodology suffers from some shortcomings. First, perceptual sociolinguistic meaning emerges out of a speaker's use of various variants and the way they combine them (Campbell-Kibler 2011). A single item, such as an added emoticon, cannot capture the complexity of this process. In order to obtain analyzable and interpretable results in a dataset of this size, however, this simplified approach was chosen. Two issues need to be considered regarding sample size and the data quality.

First, the sample size for treatment stimuli might not be large enough to identify all relevant differences. A power analysis – power analysis uses statistical algorithms to help researchers estimate the sample size needed for their study to achieve desired level of statistical power (Cohen 1992) – using the R library *pwr* (Champely et al. 2017) – suggests that 100 participants per stimulus ought to be sufficient to pick up on large effect sizes (defined as Cohen’s $w > 0.5$). Anything below that threshold, however, might go unnoticed.

Second, little quality control is possible on Amazon Mechanical Turk. Participants could easily cheat or lie on the survey. Participants’ input was only checked for basic plausibility (such as: does the place of residence really exist?). One indicator of productive participation, however, is the substantial number of comments left in the “Further comments” box. Some comments engaged with study design (“I did not understand question x”; “This ad made me uncomfortable”), others volunteered information (“I think it’s a man because ...”). The time spent on the survey by participants (average around 3 minutes) also suggests that participants did not just click random buttons.

4.3. RESULTS

The results were collected on and then downloaded from Qualtric’s web survey service. All items that did not contain a response to the “What is the author’s gender” question were excluded. All other empty fields were coded as “NA”. Any identifying participant information such as the Mechanical Turk worker ID was removed.

The final dataset included 891 participants (473 female, 411 male, 3 non-traditional gender, 4 NA). The mean time to completion was 209 seconds (median 174 seconds). The mean self-reported participant age was 35.7 years (range 18 to 74). To test

whether participant populations were similar across stimuli, chi-squared tests of independence were run on participant gender, sexual orientation, ethnicity and age for each treatment stimulus, comparing the population demographics to the control stimulus. A significant difference ($p < 0.05$) in gender makeup was found for the prosody and the punctuation stimuli, both of which contained a significantly higher number of male participants than the control stimulus. The punctuation stimulus was also exposed to a population significantly different in ethnic composition from the control, with a higher number of participants identifying as Asian. Of the completed surveys, the section of stimulus-related questions included 214 NAs (not including the 320 surveys where the author attractiveness question was omitted). The participant demographic section included 62 invalid entries.

The two sections to follow summarize the initial results regarding the control stimulus (4.3.1.) and the various treatment stimuli (4.3.2.). These are followed by a discussion of constructing a genre-consistent dataset and discussion of gender and author attribute perception.

4.3.1 Results control stimulus

Of the 200 participants in the control stimulus evaluation task, 121 perceived the author of the control stimulus to be male, 81 chose female. (Participant gender did not make a difference here; both groups voted 59 percent male in this survey). This means that the control stimulus was not perceived as completely gender-ambiguous by participants. The difference from a chance distribution is statistically significant at $p < 0.05$ ($chi\text{-squared} = 7.57$, degrees of freedom = 1, $p = 0.006$). This indicates that the control stimulus is not gender-neutral: participants are significantly more likely to

attribute it to a male author. While not required for the statistical analysis, a perceptually truly gender-balanced stimulus would have been preferable. Several comments left by participants, however, do comment on the opacity of the text. Writing in the “Further comments” box, one participant points out that “[t]he text is very generic. Could actually be used for male/female, homosexual/heterosexual, and almost any race.” Another participant writes that “I don’t really have any suspicion about race or gender from was written [sic], I’m just giving my gut reaction.”

4.3.2. Results treatment stimuli

To test for significant differences to the control stimulus, chi-squared tests of independence (Crawley 2007:222–23) were run against each treatment stimulus. Significant differences in gender perception to the control were found for the prosody and the clippings stimulus.

4.3.3. Controlling for genre

Before analyzing the data in more detail, it is important to consider text genre as a potential confounding factor. Inspection of the dataset and participant comments suggest that not all participants read the stimulus as a dating ad. Quite a few participants, for example, perceived the author to be a heterosexual female writing to another female – a scenario not possible in a dating ad focused on a romantic relationship. Or consider for instance this comment from the control stimulus: “it seems like it’s a young-ish woman looking for friends..”. That is to say, some participants did not read the stimulus text as coming from a writer looking for a romantic partner, but assume that the writer is looking for a friend or platonic relationship. Results from these participants might consequently not be comparable to the findings from the production study since we cannot know what

genre effects apply in each context: feature perception as well as linguistic production might differ between a “Strictly platonic” ad and a dating ad. To control for this external variable as much as possible, all participants whose replies to the questions about author gender, author sexual orientation and addressee gender indicated that they did not consider the stimulus text a dating ad were excluded. That is to say, participants who fell into either of the following two categories were removed:

- 1) Participant chose author gender “female”, author sexual orientation “heterosexual”, and addressee gender “female”; or participant chose author gender “male”, “heterosexual”, addressee “male”.
- 2) Participant chose author gender “female”, author sexual orientation “homosexual” and addressee “male”; or participant chose author gender “male”, “homosexual”, addressee “female”.

In short, this excludes any constellation where an author is perceived to address an audience that they are not potentially romantically interested in. This includes authors perceived as heterosexual men writing for another man, or authors perceived as homosexual men writing for a woman.

Removing these genre-ambiguous responses left 531 items in a genre-consistent perception dataset. 114 responses out of 201 were retained for the control stimulus, 53 (100) for the capitals stimulus, 60 (98) for the clipping stimulus, 60 (100) for the prosody stimulus, 58 (98) for punctuation, 52 (97) for the single letter stimulus. Thus, this step weakens the statistical power of our models but ensures that we apply the same strong genre criterion to the perception study as we did to the production study. However, even in this more genre-consistent dataset, it cannot be guaranteed that participants perceive the author to be looking for a romantic rather than a platonic partner. As shown below, this focusing of the dataset changed details, but not the overall trend of the results.

From now on, the discussion will focus on results from this genre-consistent dataset, which is more appropriate for the task of comparing perception results to the outcome of the production study.

Overall results for the original and the genre-consistent dataset are given in table 30 below. The distribution of individual stimuli over perceived author attributes is plotted in Figure 10 and Figure 11 below.

Response Type	#	Question	Results (results after controlling for genre)
multiple choice:	1	The author is ...	<input type="checkbox"/> male: 470 (320) <input type="checkbox"/> female: 420 (218)
binary	2	The author is ...	<input type="checkbox"/> homosexual: 61 (52) <input type="checkbox"/> heterosexual: 826 (486)
	3	The author is writing for ...	<input type="checkbox"/> a man: 399 (255) <input type="checkbox"/> a woman: 489 (283)
multiple choice: 4 items	4	I'd guess the author is ...	<input type="checkbox"/> Asian: 33 (33) <input type="checkbox"/> Black/Af.-Am.: 18 (18) <input type="checkbox"/> White: 812 (482) <input type="checkbox"/> Hisp./Latino: 29 (22)

Table 30: Questions assessing author perception in the matched guise study: response count for multiple choice questions, means for Likert questions. Results for genre-consistent dataset in parentheses.

Likert scale: 5 items	5	The author seems ...	educated mean= 2.741 (2.738)
<input type="checkbox"/> Very __	6		assertive mean = 2.754 (2.767)
<input type="checkbox"/> Somewhat __	7		sensitive mean = 2.804 (2.807)
<input type="checkbox"/> _	8		friendly mean = 1.934 (1.989)
<input type="checkbox"/> Not Very __,	9		attractive mean = 2.946 (2.964)
<input type="checkbox"/> Very un__.	10	Would you reply to this ad?	likely mean = 3.883 (3.93)
Open-ended	11	Men are...	NA
		Women are ...	
	13	Do you have any further comments? Please share!	NA

Table 30: continued.

4.3.1.1. Binary features: perceived author gender, perceived author sexual orientation, perceived addressee gender

Questions with binary response in the questionnaire include the perceived author gender, perceived addressee gender, and the author's perceived sexual orientation.

Within those, two larger trends are apparent across stimuli: one is a tendency to regard the author as male. While some individual stimuli are perceived by a majority as female-authored, 59 percent of all participants classified the stimulus they were working on as male-authored. Results for the question regarding the author's sexual orientation are even more one-sided: 91 percent of respondents chose "heterosexual". (This ensures that the other binary, perceived addressee gender, is essentially the mirror image of the perceived author gender). After discussing high-level trends in our dataset, it remains to analyze individual treatment stimuli and compare them to the control stimulus.

When looking at differences between the control stimulus and individual treatment stimuli, perception of author gender is significantly different at the $p < 0.05$ level between the control and the emoticon stimulus (*chi squared* = 7.911, degrees of freedom = 1, $p = 0.0049$) and the prosody stimulus (*chi squared* = 13.953, degrees of freedom = 1, $p = 0.0001$)

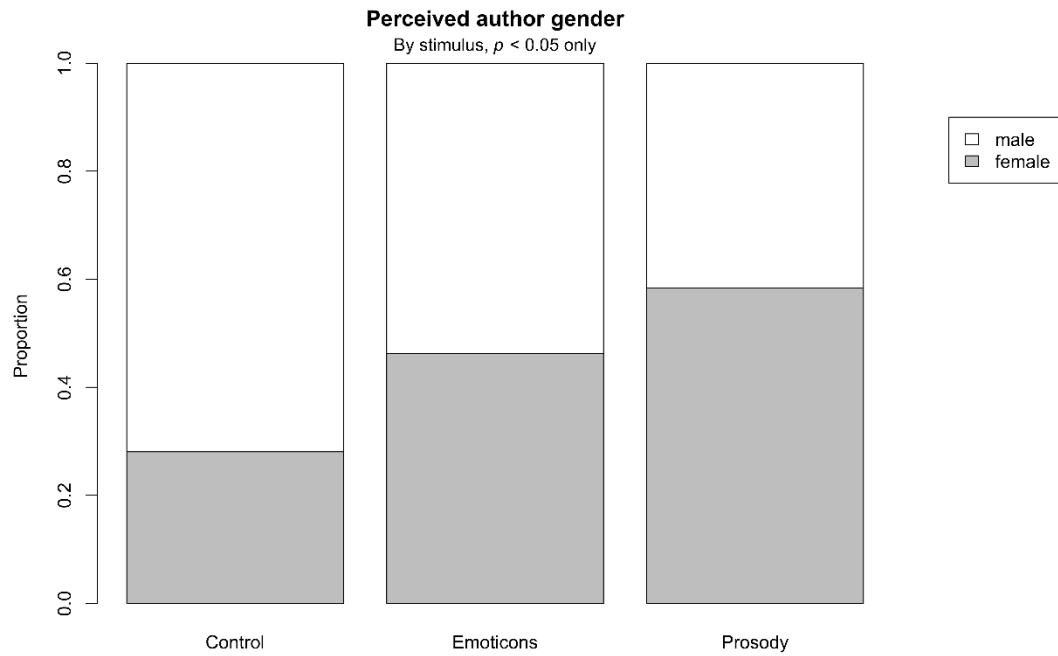


Figure 9: Perceived author gender by stimulus: responses to emoticons stimulus and prosody stimulus, compared to control stimulus.

As noted above, perceived addressee gender is merely a function of perceived author gender, with only 47 items in the dataset listing perceived author sexual orientation as homosexual. It thus simply mirrors the significance values for perceived author gender.

Author perception	Control	Capitals	Clippings	Emoticon	Prosody	Punctuation	Single Letter
Author male	72 %	64 %	60 %	54 % *	42 % *	66 %	58 %
Author female	28 %	36 %	40 %	46 % *	58 % *	34 %	42 %
Author heterosexual	90 %	91 %	88 %	97 %	97 %	87 %	96 %
Author homosexual	10 %	9 %	12 %	3 %	3 %	12 %	4 %

Table 31: Perceived author gender in the matched guise study, percentages by stimulus. * indicates differences to control stimulus significant at the $p < 0.05$ level.

The clipping stimulus, however, in the genre-consistent data can no longer be considered significantly different from the control stimulus regarding perceived author gender ($chi\ squared = 1.727423$, degrees of freedom = 1, $p = 0.1887$).

Collating perceived author gender and addressee gender results in Craigslist-like categories (m4w, w4m, m4m, w4w) parallel to the ones inspected in chapter 2. Since there is so few data on non-heterosexual authors, however, the results are not very interpretable.

	Control	Capitals	Clippings	Emoticons	Punctuation	Prosody	Single Letters
m4m	8.77 %	7.54 %	1.88 %	11.94 %	12.06 %	3.33 %	1.92
m4w	63.15 %	56.60 %	58.49 %	41.79 %	53.44 %	38.33 %	55.76
w4w	0.87 %	1.88 %	1.88 %	0	0	0	1.92
w4m	27.19 %	33.96 %	37.73 %	46.26 %	34.48 %	58.33 %	40.38

Table 32: Perceived Craigslist category in the matched guise study, percentages by stimulus.

4.3.2. Participant meta-commentary on gender-linked features

Participants' comments can add some further perspective on the findings that prosodic features and emoticons were perceived to be more likely coming from a female author. The comments discussed here were extracted from the "Men are ..." and "Women are ..." prompts in the questionnaire.

Of the 98 participants in the emoticon questionnaire, 94 participants filled out both these boxes. Of the 100 participants exposed to the prosody stimulus, 91 responded to the "Women are ..." prompt, and 90 to the "Men are ..." prompt. Most of the replies consist of very brief entries. Popular options include "from Mars" and "from Venus", "male" and "female", "humans", or evaluative adjectives such as "awesome", "cool", "dumb". A number of participants referred back to the perceived author characteristics questions they had answered earlier, for instance entering "assertive". It also needs to be noted that several participants "did not understand" the two questions or felt that they were "too open-ended" (these comments are from the "Further comments" section of other stimuli). One participant noted that they just "wrote whatever I think you wanted to hear".

4.3.2.1. Meta-commentary prosody stimulus

Participants left 181 comments in the "Men are ..." and "Women are ..." boxes after rating the prosody stimulus. These comments suggest that some participants picked up on the specific linguistic features the stimulus was testing for (*haha* and *sooo*). Additionally, they shed light on some of the participants' ideas about gender. For the prosody stimulus, comments from the prompt "Women are ..." include

Women are ...

- 1) *more likely to use laughing and draw out their words in text than men.* [41]
- 2) *more chatty, like this author, and haha looking for something* [100]

3) *Women seem more playful in their writing style I think.* [82]

4) *More prone to inputting [sic] emotion into texts.* [97]

Comments 1) and 2) explicitly single out the prosodic features in question, the items *haha* and *sooo*. Comments 3) and 4) could be understood to pick up on the same feature.

Comments from the “Men are ...” prompt of the prosody questionnaire include:

1) *more dominant in their writing and not as much emotion. Often more direct.*

[94]

2) *likely to respond to this ad because it seems flirty and interesting.* [84]

3) *This writing style doesn't fit a man.* [82]

4) *Less flirty than this.* [97]

Note that comments 3) and 4) from both sections are written by the same participants in the “Men are ...” / “Women are ...” boxes respectively.

4.3.2.2. Meta-commentary emoticon stimulus

Participants left 175 comments in the “Men are ...” and “Women are ...” boxes after rating the emoticon stimulus. These comments suggest that participants picked up on the specific linguistic features (:)) the stimulus was testing for. Additionally, the comments shed light on participants’ ideas about gender. The commentary explicitly discussing linguistic aspects of the emoticon stimulus questionnaire includes:

Women are ...

1) *usually more flirty and more likely to use smiley faces.* [emoticons 82]

2) *Much more likely to come across as approachable and very friendly, sometimes through use of emotes but also through a generally friendly way of speaking.* [...]

[emoticons 57]

- 3) [...] *Generally from my friends I've just noticed women use those smileys more than men leading to my guess.* [emoticons 24]

Men are ...

- 1) *Men are less likely to use emotes as much as this person did.* [emoticons 57]
- 2) *usually more assertive and less likely to use smiley faces.* [emoticons 82]

Note that in this case all comments regarding men cited above come from participants that also commented on women. Obviously, not all comments align with the results as clearly as the ones above. It is noteworthy, however, that none of the participants making the argument that the stimuli discussed above were written by a male author referred to linguistic features in their comments. The following comments illustrate this point:

Men are...

- 1) *more likely to want to go do macho things, which usually does not include going to the beach and drinks. I could see them more wanting to go to a sports game or go camping.* [prosody 49]
- 2) *generally less willing to talk about interests* [prosody 91]

That is to say, these comments are discussing content, rather than style, of the ad. We find similar comments about women as well:

Women are ...

- 1) *in need of more social interaction and therefore more open to meeting new people. I loved those times when I was single and went out for drinks with my girls* [prosody 49]
- 2) *More interested in hanging out in a group like this writer* [prosody 18]

4.3.3. Dynamics of perceived author gender and perceived author attributes

In a second step, the gender results discussed in section 4.3.2. are considered in the context of results regarding the other attributes such as friendliness, education, and assertiveness queried in the survey. Figure 10 and Figure 11 below show the results for the entire dataset and the genre-consistent dataset, where means for each stimulus are plotted in comparison to the mean of the control stimulus (plotted as $y = 0$). The y-axis thus represents the distance of the treatment stimulus' mean to the control stimulus' mean. To create this graph, participants' responses to the Likert scale questions were converted to numbers after ordering them from "Very __" to "Very un__". For each question, the first option (such as "Very educated") was converted into the number 1, the second option ("Educated") into the number 2, and so on. (For example: Ten "Very educated" ratings lead to a score of $10 * 1 = 10$ and thus a mean of 1; ten "very uneducated" lead to $10 * 5 = 50$ and a mean of 5). In R, responses were imported as factors, then ordered as described, and converted to numbers.

In the plot below, scores above zero indicate that on average participants gave a rating more positive than participants did on the control stimulus; scores below zero indicate that the average score was lower than the control mean. According to Figure 10, for example, the author of the prosody stimulus was on average perceived as friendlier than the author of the control stimulus; the punctuation stimulus, on the other hand, was perceived as less friendly than the base line.

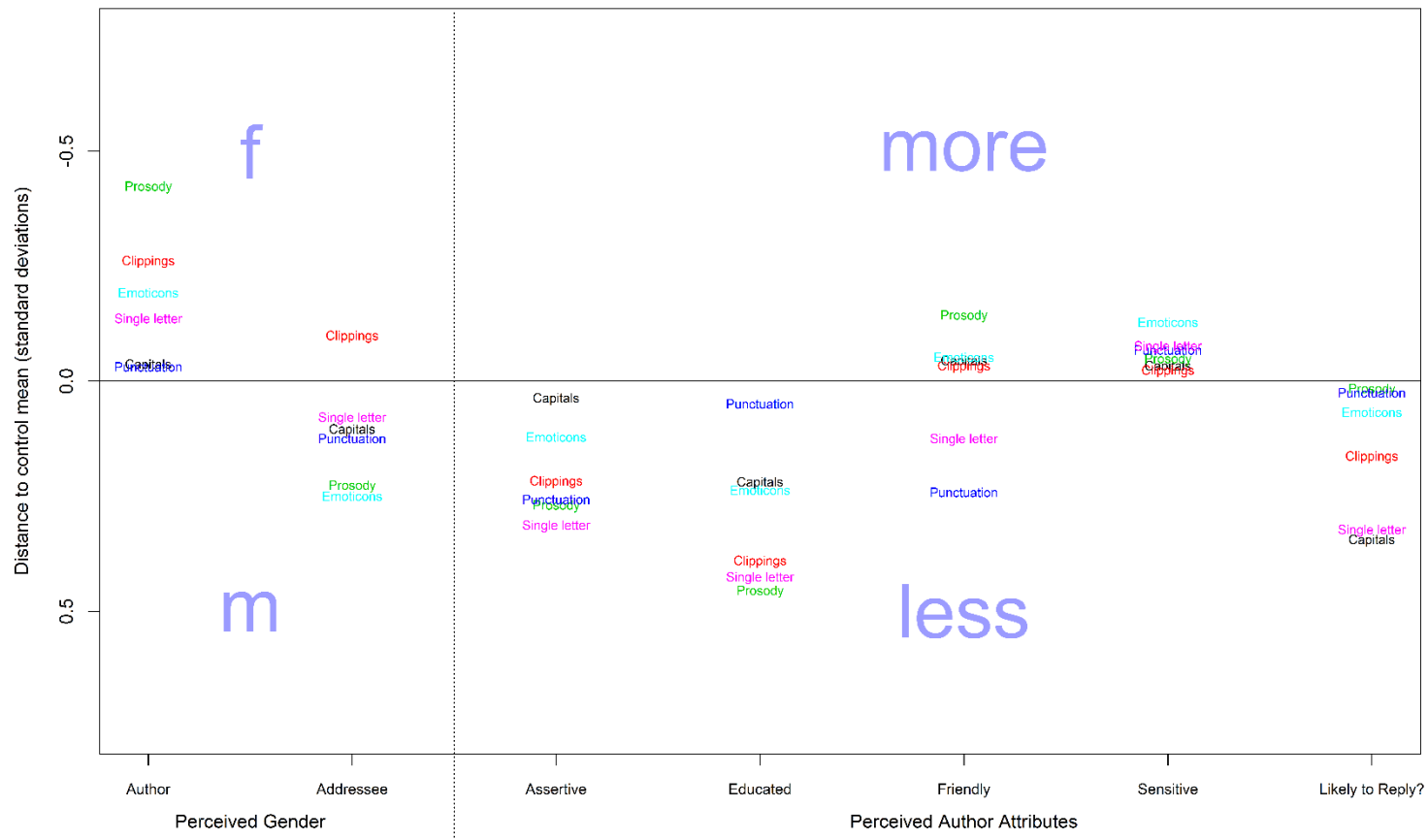


Figure 10: Full dataset: Means of perceived author characteristics, relative to control stimulus.



Figure 11: Genre-consistent dataset: Perceived author characteristics, relative to control stimulus.

Some basic findings evident in Figure 11 include: Compared to the control stimulus, no stimulus was perceived as more likely to be written by a male author. Statistically significant differences in this regard, as described above, exist only for the emoticon and prosody stimulus. Non-significant differences based on visual inspection include: all stimuli authors, except for the emoticon stimulus, were perceived as less assertive than the control stimulus. Results for perceived author friendliness and perceived author sensitivity cluster around the control mean. However, all authors except for the punctuation stimulus were perceived as less educated than the control stimulus author. Two of them differ from the control in a statistically significant way: the clippings stimulus (*chi squared* = 14.367, degrees of freedom = 4, *p* = 0.006) and the prosody stimulus (*chi squared* = 10.36771, degrees of freedom = 4, *p* = 0.034). These results for perceived author education are shown in Figure 12 below.

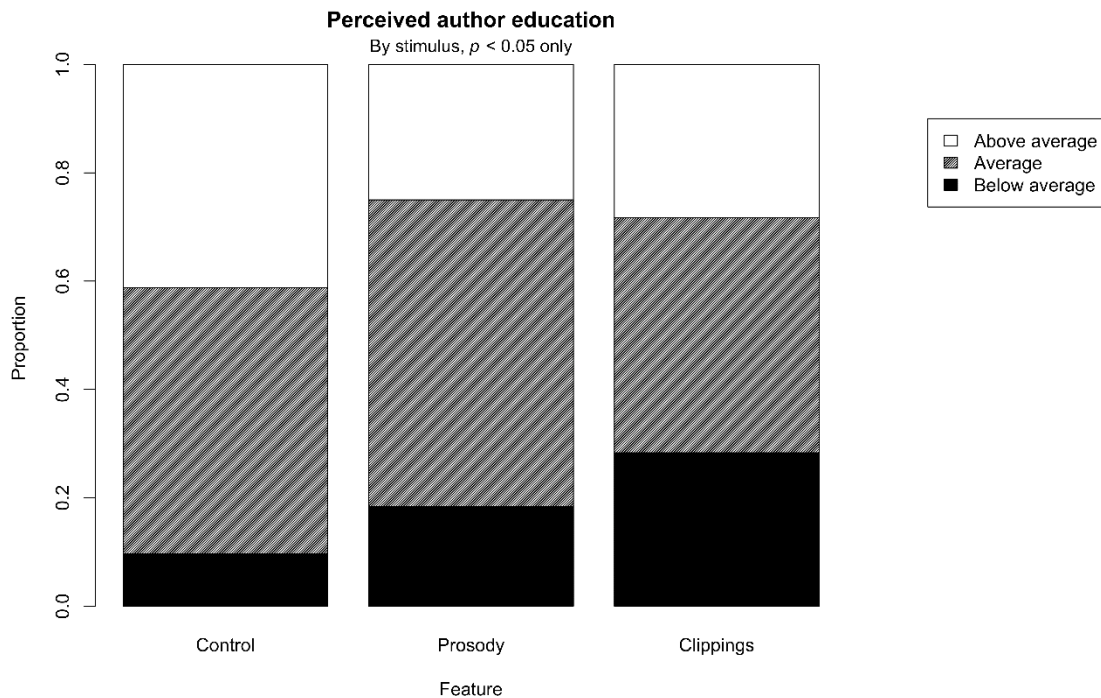


Figure 12: Perceived author education by stimulus: responses to clipping stimulus and prosody stimulus, compared to control stimulus.

Visually, the punctuation stimulus is somewhat of an outlier in general: it is the only stimulus that is rated above the control stimulus in education, it is rated noticeably lower than the control stimulus on friendliness, and is the only item to be above the control group in participants' willingness to reply to the ad.

Regarding the two significantly gendered features prosody and emoticons, several noteworthy patterns emerge. Both stimuli, in addition to being perceived as more likely to be written by a female author, are also above all other stimuli in perceived friendliness. Both are perceived as written by less educated authors than the control stimulus, the prosody stimulus significantly so. No even visually consistent patterns can be identified for correlations with perceived assertiveness or perceived sensitivity. It must be kept in

mind that only on perceived gender and perceived education on the prosody stimulus reach statistical significance.

4.3.4. SUMMARY

We can thus summarize the results of the study regarding each of the hypotheses initially formulated in 3.1. as follows:

H 1 Certain writing styles and linguistic variants are perceived as masculine or feminine.

Result When compared to the control stimulus, two of the seven features studied affected the perceived gender of the author: prosodic items and emoticons. Both stimuli were significantly more likely to be perceived as written by a woman.

H 2 Gendered features will have other social meanings attached to them.

Result The prosody stimulus was perceived significantly lower on perceived author education than the control stimulus. To a lesser extent, both the prosody and the emoticon stimulus were both rated higher than the control stimulus on perceived author sensitivity. The clipping stimulus, which did not appear to have a gendered meaning, was also significantly below the control mean in perceived author education.

H 3 Gender perception will vary by the study participants' own gender, age, and other social characteristics.

Result While participant demographics were controlled for when analyzing responses to the survey questions, correlations between participant demographics and author perception were not analyzed due to time constraints.

H 4 Perception of sexual orientation will interact with gender perception.

Result Since the author's sexual orientation was overwhelmingly perceived as heterosexual, H4 could not be addressed.

Chapter 5: Conclusion

This chapter presents a synthesis of the results and discussions in chapter 3, the production study, and chapter 4, the perception study. To do so, it will consider the results in their social context; establish relevant linguistic features; and present an attempt at explaining the use and social meaning of these features in a systematic way. In line with the methodological considerations outlined in section 2.7., gendered features will be identified by looking for gender-linked patterns in both production and perception. Results from both the production and the perception study will be brought to bear upon the discussion of the linguistic indexicality of these features that follows in section 5.4.. This analysis of indexical meaning pays close attention to the social and linguistic context by considering the text genre and by tying linguistic features to American gender ideologies established in the perception study and previous research on gender stereotypes.

5.1. SUMMARY: RESULTS OF CHAPTERS 3 AND 4

Below, the hypotheses tested in each chapter with the respective findings are reproduced as a starting point for discussion.

5.1.1. Chapter 3: production

Below are the hypotheses and findings from the analysis of language production in the dating ad corpus, which are also illustrated in Figure 12.

H 1 There is variation by binary author gender (male/female) in the data.

Result Author gender is a weak predictor in this dataset. An explanation in terms of binary author gender only obscures possibly meaningful patterns.

H 2 There is variation by binary addressee gender (male/female) in the data.

- Result** Similarly, addressee gender is a weak predictor in this dataset. None of the features pattern strongly by addressee gender.
- H 3** There is variation by the author's and addressee's sexual orientation (binary: heterosexual/homosexual) in the data.
- Result** H3 is true for several features. Several features where m4m and w4w authors lead in feature use can be interpreted this way.
- H 4** There is variation by an interaction of the above in the data.
- Result** Hypothesis 4 is supported by the results of the production study in non-straightforward ways. Feature frequencies vary quite strongly by author – addressee dyads: for instance, m4m authors differ linguistically from m4w authors. This suggests a linguistic accommodation effect or variation by sexual orientation. One general pattern that emerges out of the data is that men writing to other men (m4m) tend to use e-grammar features other than emoticons and prosodic items more frequently than the other dyads. When writing for women (m4w), on the other hand, male authors are below the overall mean in the use of the very same features. Similarly, women writing for other women (w4w) tend to use e-grammar features more frequently than when writing for men. This applies to abbreviations, capitalized words, and clippings. The two features emoticons and prosody exhibit a strikingly different pattern: here, male authors writing for a male audience are on average the least frequent users, with women writing for other women being most prolific and w4m and m4w authors converging around the mean. Especially for these two items, an interaction in the form of a linguistic accommodation effect is suggested by the data.

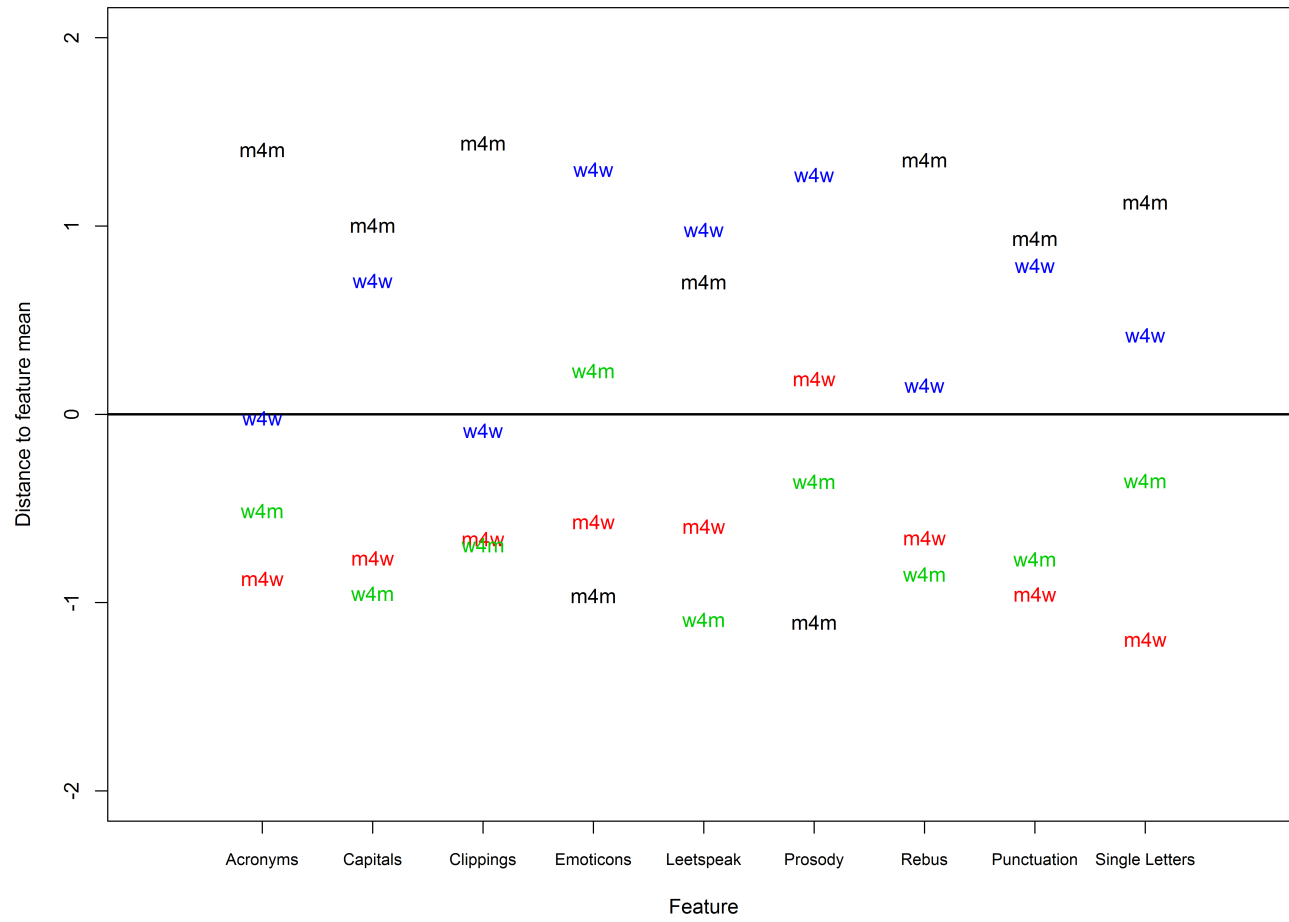


Figure 13: Relative feature frequency by category.

5.1.2. Chapter 4: perception

Below are the hypotheses and findings from the analysis of language perception, which are also illustrated in Figure 13.

H 1 Certain writing styles and linguistic variants are perceived as masculine or feminine.

Result When compared to the control stimulus, two of the seven features studied affected the perceived gender of the author: prosodic items and emoticons. Both stimuli were more likely to be perceived as written by a woman.

H 2 Gendered features will have other social meanings (e.g. traits such as assertiveness) attached to them.

Result The prosody stimulus affected the perceived education of the author. It was perceived as significantly “less educated” by participants. Visual inspection, though not statistical significance testing, suggests that this pattern holds true for large parts of the dataset: perception of author gender as female co-occurs with lower ratings on perceived author education. To a lesser extent, perception of female author gender also tends to lead to higher ratings in perceived author friendliness.

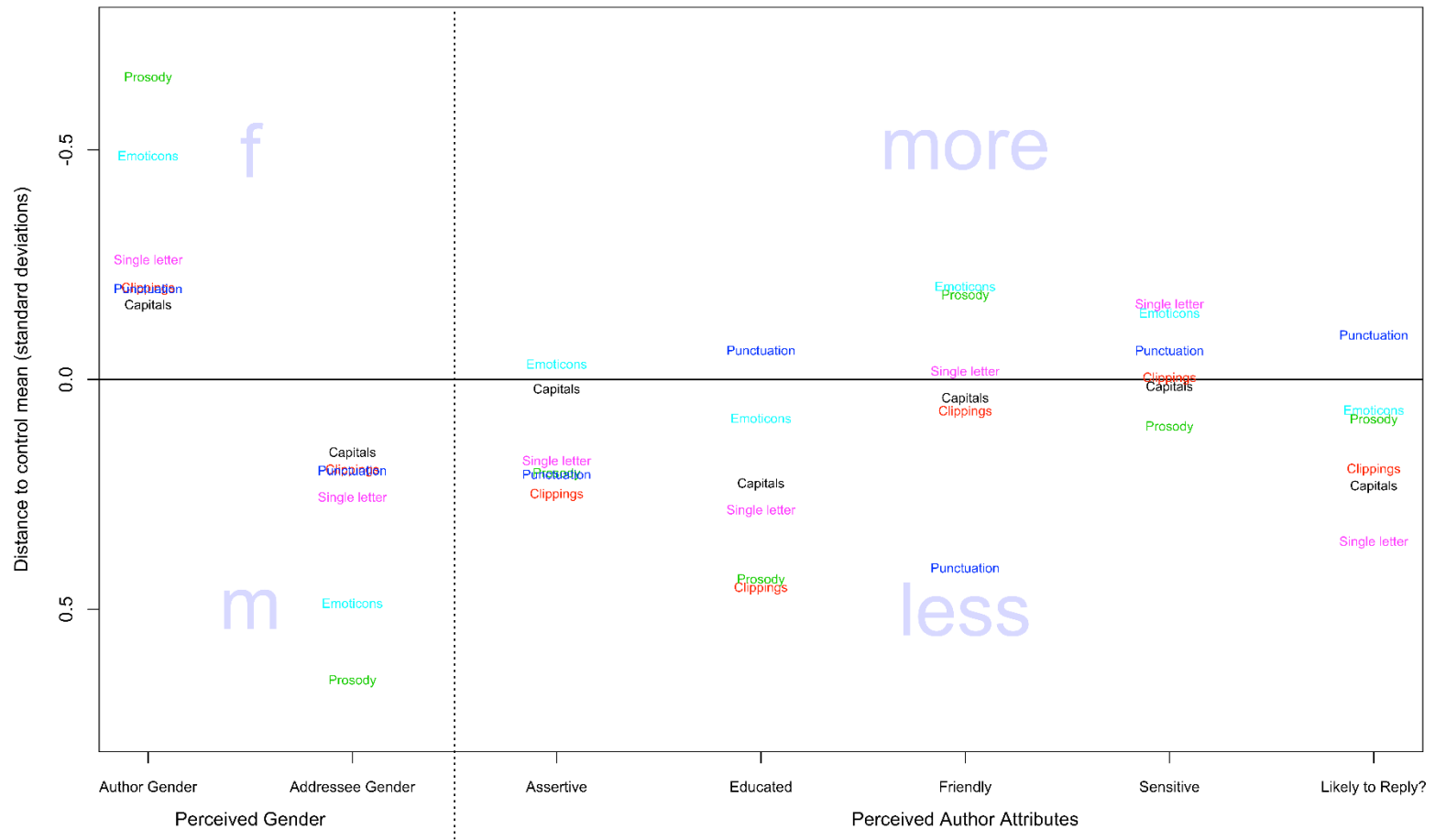


Figure 14: Genre-consistent dataset: Perceived author characteristics, relative to control stimulus.

5.2. INTRODUCTION

The goal of this final chapter is to place the findings of the production and the perception study in their linguistic and socio-cultural context. This will allow us to get to the heart of the matter: how is language used to produce gender differentiation, and what social meanings are associated with gendered features?

The studies presented in chapters 3 and 4 constitute attempts at illuminating the two sides of indexicality: first, the production study investigated writers' attempts at constructing a certain gender identity through linguistic means; then, the reader's perception of these linguistic means was analyzed in the perception study. As noted before, the rationale for this approach is that both writer and reader are part of the meaning-making process (Agha 2007; Campbell-Kibler 2010; Walker et al. 2014). To integrate the results from the two studies and explore the indexicality of e-grammar features found to be gender-relevant, a close analysis of the context writer and reader are operating in is necessary. Thus, the identification of gendered features in section 5.3. is followed by a discussion of genre conventions, gender ideologies, and addressee effects. These constitute the regulatory framework (Butler 1999) of gender performance in this context and are addressed in sections 5.4.1. to 5.4.3. below. The chapter concludes with a discussion of how social context interacts with and influences the indexicality of e-grammar features in the present setting.

5.3. INTEGRATING PRODUCTION AND PERCEPTION

5.3.1. Points of divergence

The most striking incongruity between the production and perception study is the conception of gendered groups. The author-addressee dyad that most appropriately described linguistic variation in the production study was not reflected in the perception

study: this distinction was not at all apparent in the participants' perception and evaluation of linguistic features. Instead, all participants assumed a binary model in which a heterosexual writer addresses the other gender, thus effectively pre-supposing the addressee based on author perception. There was thus little to no data regarding language use for the categories m4m and w4w.

In short, while the analysis of production data signaled a move away from the traditional gender binary, the participants of the perception study approached the evaluation task with exactly this binary in mind. This points to an intersection of linguistic practice and social ideologies, in this case the premise that the writer is heterosexual, or the heteronormative assumption (Kiesling 2009). This incongruity suggests that the nuances of linguistic practice in this case are eclipsed in perception by a more essentialist social ideology. The differences in performance for different audiences that are observed in production seem not to all that meaningful for gender perception. In this case, an ideological system centered on binary gender and heteronormativity does not leave room for linguistic indexicalities beyond these assumptions. This again brings home the point that the social meaning of linguistic features cannot be based on analysis of linguistic production, or "abstract patterns" (Campbell-Kibler 2010), alone, but needs to take into account the social ideology of the audience encountering the linguistic form.

5.3.2. Points of convergence

The two studies did converge, however, on two gendered e-grammar features: a text is more likely to be perceived as female-authored if it contains prosodic items such as *haha* or emoticons such as :). Both features are also used more frequently by female authors in the production study. A positive correlation between female authors and emoticon use in online writing has also been reported by Rao et al. (2010), Burger et al.

(2011) and Bamman et al. (2014). (No studies of prosodic items as defined here could be located). We can thus postulate that these two features are correlated with female writers in both production and perception. No such argument is possible for any of the other e-grammar features.

A closer look at the distribution of emoticons and prosodic items over the four categories m4m, w4m, w4w and m4w, however, complicates this picture. While it is true that female authors overall use these features more frequently than men do, results from the production study indicate that women use them a lot when writing to other women (w4w), but less when writing for men (w4m). Men, on the other hand, barely use them when addressing other men (m4m) but tend to employ them more frequently when addressing women (m4w). Indeed, feature frequencies are not significantly different between the m4w and w4m categories. Thus, we might hypothesize an audience effect: use frequency of these gendered features is not only dependent on author gender, but also addressee gender.

It is tempting to read these results as women toning down their “female way of writing” when writing ads for men, and men acting “more feminine” when addressing women (cf. Hogg 1985). However, the discussion below is intended to illustrate that there is no one “female way of writing”. Rather, writing like a woman, which we can understand as part of performing female gender, depends on factors like the audience someone is writing for. Features perceptually associated with female writers, such as prosodic items and emoticons, are important in this process since they can be used to do the accommodation work described above. There is nothing, however, that makes them inherently female. The process of how they come to be perceived as “female” and enter gendered variation will be addressed in the next section: why are some features linked to female authors, in production as well as perception? How does a linguistic item get to be

“feminine”? This is the realm of the study of indexicality, which the next sections will focus on. It presents a discussion of the indexicality of emoticons and prosodic items, which will explore the social meanings of these features under consideration of the present context.

5.4. INDEXICALITY OF EMOTICONS AND PROSODY

In sociolinguistics, indexicality refers to the potential of linguistic items to acquire social meanings. Indexical items, according to Ochs (1990:288) “vary across contexts and hence index (point to) contexts when used”. Indexicality in gender-linked variation specifically is discussed in Ochs (1992), who argues that indexing gender is an indirect process. Linguistic variants, rather than signaling a certain gender identity directly, are first linked to other social attributes, such as politeness. Based on this original index, they can acquire a gender-linked meaning. If, for instance, gender stereotypes hold that women are more polite than men, the politeness-indexing feature can become a femininity-indexing feature. Thus, their meaning is determined by what Butler (1999) in her more general definition of gender calls the “regulatory frame”:

Gender is the repeated stylization of the body, a set of repeated acts within a highly regulatory frame that congeal over time to produce the appearance of substance, a natural sort of being.

(Butler 1999:43–4)

In the present study, this regulatory frame thus consists of gender ideologies on the one hand, discussed in section 5.4.3.; other parts of the regulatory frame that will be discussed below are the genre conventions of dating ads (section 5.4.1.) and the impact of the audience (section 5.4.2.). All these discussions will draw on insights from the production and perception study as well as external research and will help determine which indexical

meanings are activated in the present context. This approach builds on the notion of the indexical field, developed by Eckert (2008a). Eckert's (2008a) paper expands on the concept of the "indexical order" (Silverstein 2003), the idea that linguistic indexicalities emerge out of previous social meanings. Ochs's (1992) account of gendered indexicalities building on previous linguistic indexes can be read as an instance of indexical ordering, where gender always is an $n + 1$ th order. The indexical field Eckert proposes based on this research consists of a collection of the meanings a given variable can index depending on its linguistic and social context. Eckert (2008a:469) argues that these indexical meanings are highly context-dependent. In order to understand which indexicalities of emoticons and prosodic items are activated in the present study, we therefore need to look at the linguistic and social context they are occurring in. This is what the sections to follow are attempting to achieve.

5.4.1. Genre conventions

In studies of gender in writing that do control for genre, genre emerges as a stronger predictor than gender (Rubin & Greene 1992; Janssen & Murachver 2004; Herring & John C. Paolillo 2006). The pitfalls of ignoring genre in gender studies are for instance pointed out in Herring & Paolillo's (2006) refutation of Koppel et al. (2002), a study that conflated gender markers with genre markers by looking at linguistic features across various genres. The present study controls for genre effects by focusing on one genre: online dating ads. However, several parts of the study suggest that this does not mean that the genre constitutes a non-gendered context. Rather, it appears that dating ads are perceived as a male genre, in the same way that for instance gossip is often considered a female genre (Coates 1989), as is diary writing (Heilbrun & Politt 2008), while scientific writing can be considered a masculine genre (Tillery 2005). Two aspects

of our datasets support the argument for dating ads as a similarly gendered genre: first, comments from the perception study, analyzed in light of previous research on dating conventions, and second, the overall makeup of the dating ads corpus.

In the perception study, several participants commented on their rationale for choosing “male” as the author gender: “Just asking to randomly hang out with anyone available strikes me as such a guy thing to do” (from a response to the control stimulus). Other participants note that men are “more likely to initiate hanging out with the opposite sex and rather bold when they think they’re attractive”, or that they are “more assertive than women when it comes to dating”.

These comments reflect the broader reality of American dating conventions. Studies of “dating scripts”, that is assumptions about who does what in the dating context, consistently find that in the dating process, the male partner is expected to be proactive, while the female partner is expected to be reactive (Rose & Frieze 1993; Laner & Ventrone 2000; Serewicz & Gale 2008). This entails, as shown for instance in Laner & Ventrone (2000:493), that Americans expect the male partner to make the first move in a dating situation: “the woman’s role on the first date is a reactive one”. In a review of dating script studies conducted since the 1970s, Eaton & Rose (2011:853), find that even today, studies find young Americans “reproducing the hypothetical scripts generated by young adults [...] 20 years earlier”.

Note that the research cited above documents the behavior of heterosexual college students; thus, these findings might not generalize across populations. However, little research has been done outside of this context. This ties in with the discussion of potential sub-genres in section 3.3.3.2.; it is possible that the ideologies and stereotypes discussed here only apply to heteronormative dating and that the social ideologies and dating conventions are different for m4m and w4w writers. However, since these writers are still

operating in the context of the mainstream American gender system, they cannot separate themselves completely from this ideological context. While it is likely that they are creatively engaging with or subverting ideological norms and gendered stereotypes (Thorne & Coupland 1998), there is no way to discuss this in more detail given the lack of relevant data in the perception study as well as external research.

The cultural convention of male agency in dating has also been illustrated by studies of dating manuals, that is handbooks telling people how to date successfully. Laner & Ventrone (2000) as well as Eaton & Rose (2011) show that the advice given in such books typically reinforces the ideas of male agency present in dating scripts. Analyzing the ten best-selling dating manuals from online bookstore *Amazon.com*, Eaton & Rose (2011:845) conclude that “this select set of current popular books generally endorsed traditional feminine passivity and masculine agency in the dating context”. Eaton & Rose (2011:845) cite passages from these books that advise women “not to object to his [i.e. their male date’s] plans unless you really have to” and men to act like a “natural born aggressor”. Or consider the chapter headings from *The Rules: Time-tested secrets to capture the heart of Mr. Right* (Fein & Schneider 1995), a dating guide for women. Its first four chapters are titled

1. *Don't Talk to a Man First (and Don't Ask Him to Dance)*
2. *Don't Stare at Men or Talk Too Much*
3. *Don't Meet Him Halfway or Go Dutch on a Date*
4. *Don't Call Him and Rarely Return His Calls*

[...]

While there is little empirical research on dating conventions for online dating or interaction on mobile dating apps (except for Long (2010), who argues that dates are initiated by the match-making algorithm rather than a person in this context), their users

seem still aware of these conventions. Consider *Bumble*, a smartphone dating app that claims to be “Changing the rules of the game” (Bumble 2017) and has been downloaded several million times. The user guide explains that on Bumble, the woman always makes the first move, in an attempt to “counter the age-old and often outdated ‘guys always have to make the first move’ idea!” (Bumble 2017). This suggests that these dating conventions are present and being challenged at the same time.

Applying the logic of the dating scripts described in the research cited above to our present dataset, we would then expect that in a heteronormative dating scenario, the man is the writer, the woman the addressee of such ads. Under this assumption, it makes sense that the m4w category, which covers exactly this scenario, is by far the biggest in the dating ads corpus, making up 47 percent of ads in the entire dataset. Under this assumption, it also makes sense that the default pick of participants in the perception study is to pick “The author is male”: 120 out of 200 participants perceived the control stimulus, which did not contain any e-grammar features, as male-authored. This difference is significant at the $p < 0.05$ level ($chi\text{-squared} = 7.57$, degrees of freedom = 1, $p = 0.006$).

5.4.2. Addressee effects

The hypothesized addressee effect in the data, already outlined in section 5.3.2., can be tested by applying Bell’s (1984) audience design model (the slight reworking in Bell (2001) has no consequences for its application here) which attempts to explain stylistic variation in language use. That is to say, it focuses on intra-speaker, rather than inter-speaker variation. Bell’s (1984:145) central argument is that “[s]tyle is essentially speakers’ response to their audience”. (Note that this understanding of style was in competition with Labov’s (1966) conception of style as attention paid to speech). Audience

design thus models the process of a speaker accommodating to the speech style of their audience. Bell (1984:159) introduces roles such as auditor and eavesdropper in his multi-tiered model of a speaker's audience. Members of each of these audience groups can impact a speaker's stylistic choices, with the direct addressee exerting the biggest influence on a speaker's linguistic performance. Bell (1984:167) summarizes audience design as follows:

[A] sociolinguistic variable which is differentiated by certain speaker characteristics (e.g., by class or gender or age) tends to be differentiated in speech to addressees with those same characteristics. That is, if an old person uses a given linguistic variable differently than a young person, then individuals will use that variable differently when speaking to an old person than to a young person (cf. Helfrich 1979) - and, *mutatis mutandis*, for gender, race, and so on. Insofar as women speak differently than men, they will be spoken to differently than men.

(Bell 1984:167)

Thus, in this model, stylistic variation develops out of social variation: a variable has to be "differentiated by speaker characteristics", that is to say have social correlates, for a speaker to be able to use it stylistically in the way Bell describes. Audience design was developed with face-to-face spoken interaction in mind, but has in the meantime been applied widely to settings from television ads (Bell 1992) to historical texts (Nevalainen & Raumolin-Brunberg 2016). Several studies working within the closely related communication accommodation theory framework (Giles, Coupland & Coupland 1991) have investigated linguistic accommodation between genders, with somewhat inconsistent results. Some studies find that women are more accommodating in general, others find varying accommodation levels across linguistic features while still others find no accommodation effects whatsoever.

The first finding is evidence by Namy et al. (2002), who had 66 college students rate the amount of accommodation of 16 speakers to four recordings of male and female speakers. Namy et al. (2002:422) found that women accommodate more than men do, and that women are more likely to pick up on accommodation cues, potentially because they are more sensitive to vocal characteristics in general. Similarly, Jones et al. (1995), studying features such as interruptions, topic introductions, and back channels, conclude that in their data, collected from 100 students and employees of an Australian university, women tend to accommodate in general more than male speakers do.

Bilous & Krauss (1988), on the other hand, find an accommodation effect between female and male speakers in a language production experiment with respect to linguistic variables including interruptions, utterance length, and pauses in their study of 60 undergraduate students. Bilous & Krauss (1988) find that degree and direction of accommodation differ, with women and men both converging on utterance length and pauses, men converging towards women in back channels and laughter, and women converging in interruptions and total words uttered. Bilous & Krauss (1988:190) argue that their study shows that it is impossible to generalize over either features or gender when it comes to predicting the extent or direction of linguistic accommodation. In line with their results, Hogg (1985)'s study of language use among 24 English university students compared linguistic behavior between mixed-gender ("gender-salient") and single-gender ("gender non-salient") interactions. Discussing features such as swear words, pitch, and emotional speech, Hogg (1985:106) finds that women use "less feminine speech", such as emotional words, when interacting with men. While men do accommodate in the study, they do so to a lesser extent.

Brownlow et al. (2003) do not find any evidence for accommodation between genders in their study of language use in television interviews. Categorizing words based

on the text metric LIWC, they identify gendered styles but no effect of addressee gender. Guiller & Durndell (2007) studied gendered linguistic behavior (investigating mainly grammatical categories, such as use of pronouns) in online chat rooms among 197 Scottish college students and found strongly gendered writing styles independent of context – that is, no accommodation effect was evident in this study either.

It needs to be noted, however, that the present study differs from most of the research cited by Bell to support his theory (e.g. Labov (1966), Trudgill (1974)) as well as the communication accommodation research in several ways: first, there is no direct, face-to-face interaction in the dating ads corpus. How can the writers in the present dataset accommodate to an audience that might consist of dozens of different people, none of whom they have ever met? Bell argues that in such instances of one-to-many communication, the same constraints apply as in face-to-face interaction. Bell (1982) illustrates this point with a study of stylistic variation in the speech of radio announcers reading the news to different audiences. The audience design model assumes that in case of undifferentiated, large audiences, the speaker will accommodate to a linguistic “ideal” of the type of addressee they are trying to reach (Bell 1984:170). Several studies have applied the audience design model to online writing that way, including Tagg & Seargeant’s study (2014) of audience design in social networks among translocal communities; Androutsopoulos’ (2014) account of how multilingual Facebook users respond to and construct their audience through responsive and initiative audience design, and Rudat et al.’s (2014) research into audience design on Twitter, which shows that knowledge of their audience’s interests shapes the retweeting decisions of users.

In the present study, we can thus test Bell’s (1984:167) predictions on the two features in our set that have social meaning, prosodic items and emoticons: regarding those two female-linked features the audience design model predicts that “[i]nsofar as

women speak differently than men, they will be spoken to differently than men” (Bell 1984:167).

This prediction is borne out in the patterns present in the dataset: men on average use more prosodic items and emoticons, that is female-linked features, when writing to women. Women, conversely, tend to use fewer of these features when writing to men. Highest usage frequency is observed when women are writing to other women. We can thus understand the dynamic behind the use of emoticons and prosodic items as an instance of audience design: writers adapt their use of emoticons and prosodic items to the style of their intended audience. They tend to use more of these female-linked features when addressing a female audience.

An important caveat regarding our findings is that a single feature like an emoticon or prosodic item does not constitute a style, commonly defined as a “*set of co-occurring variables that are associated with the speaker’s own persona*” (Eckert & Rickford 2001:5, italics added). Rather, these are individual features that appear to be part of a feminine style that potentially includes a variety of other linguistics features. The stylistic accommodation described here might very well involve variation in other linguistic features, linguistic variables not tested for in the production and perception studies.

It is also important to note that in our dataset, the accommodation of women towards men and of men towards women happens to the extent that the two groups are no longer distinguishable in their average use of prosodic items and emoticons. This seems in conflict with the audience design model’s hypothesis that the speaker can “approach, but not match” the style of their addressee (Bell 1984:167). Three considerations show that this is a conflict in appearance only. First, as noted above, use of one feature does not constitute a style: matching the per-word frequency of emoticons

is not the same as matching a style. When accommodating to a female audience, male writers might for instance use emoticons in different contexts than their female audience does. They might fail to combine them with other features the way their female audience does. Most importantly, since writers are addressing an undifferentiated mass within which each addressee will differ in their personal style, the writer, using an idealized idea of their audience's style, will miss a number of the audience member's individual styles.

5.4.3. Gender ideologies

The last context constraint to be addressed here is gender ideology. The conceptualization of gender ideology here follows Eckert & McConnell-Ginet (2013:22), who define it as the “set of beliefs that govern people's participation in the gender order, and by which they explain and justify that participation” (Eckert & McConnell-Ginet 2013:22). Gender ideologies are most apparent in gender stereotypes, where a stereotype is understood as defined in Putnam (1975:147) as “a standardized description of features of the kind that are typical or ‘normal’”. Thus, gender stereotypes are shared expectations of what a man, woman or member of any other gender category is like. They answer the question: what characteristics and behavior make a subject more or less female, more or less male, etc.?

Eckert and McConnell-Ginet (2013:23) summarize common gender stereotypes as follows:

Members of any Western industrial society are likely to be able to produce the following set of oppositions: men are strong, women are weak; men are brave, women are timid; men are aggressive, women are passive; men are sex-driven, women are relationship-driven; men are impassive, women are emotional; men are rational, women are irrational; men are direct, women are indirect; men are

competitive, women are cooperative; men are practical, women are nurturing; men are rough, women are gentle.

(Eckert & McConnell-Ginet 2013:23)

(Note that this list of stereotypes was consulted when designing the perception study questionnaire).

What, then, are such gender stereotypes relevant to the US-American authors and readers in our dataset? There can of course be no comprehensive list, and attitudes will vary from person to person, but we can draw on sociological surveys as well as the perception study to identify some commonly held gender stereotypes in this population. An overview of the results from both sources is given below.

A questionnaire study by Broverman et al. (1972), later replicated in Bergen & Williams (1991), find that women and men tend to be perceived as polar opposites on 48 different characteristics. The authors stress the high level of agreement among participants (male and female respondents' attitudes show an almost perfect positive correlation) about what is considered typical of and desirable in men and women. Selected results of Broverman et al.'s (1972) survey, which was based on a sample of 154 college students, are presented in table 33 below.

In general, women scored higher on what Broverman et al. (1972:66–7) call a “warmth-expressiveness cluster” around attributes such as “gentle” and “sensitive to the feeling of others”, while participants perceive masculinity as associated with attributes in what Broverman et al. call the “competency” cluster, comprising attributes such as “ambitious” and “able to make decisions easily”. Broverman et al. (1972:60) find that their student participants also perceive these stereotypical features as desirable in each gender, and participants expect them to be present in an ideal man or ideal woman. Interviews with mental health professionals indicate that they consider these traits healthy for each

gender (Broverman et al. 1972:70). At the same time, Broverman et al. (1972) argue, masculine-associated trends are perceived more favorably by the general population, a finding also reported in Rosenkrantz (1968). A meta-analysis of seven follow-up studies to Broverman et al. (Lueptow, Garovich-Szabo & Lueptow 2001) finds that these gender stereotypes have not changed substantially over the last 30 years.

Feminine	Masculine
Competency cluster	
Not at all aggressive	Very aggressive
Not at all independent	Very independent
Very emotional*	Not at all emotional*
Very subjective	Very objective
Very easily influenced	Not at all easily influenced
Very passive	Very active
Not at all self-confident	Very self-confident
Very submissive	Very dominant
Does not hide emotion at all*	Almost always hides emotions*
Warmth-Expressiveness cluster	
Very tactful	Very blunt
Very gentle	Very rough
Doesn't use harsh language*	Uses very harsh language*
Very talkative*	Not at all talkative*
Very quiet	Very loud
Easily expresses tender feeling	Does not express tender feelings at all easily

Table 33: Masculine- and feminine-associated characteristics in the U.S. Adapted from Broverman et al (1972:70). Asterisks indicate features directly relevant to the present study.

The present study adds to these insights by testing for gender stereotypes present among the 891 participants of the linguistic perception study presented in chapter 4. The open-ended “Men are ...” and “Women are ...” questions were included in the questionnaire (see page 131 for a list of questions) for this reason. Participant responses to these questions will help us identify gender stereotypes present among the participants. While ideas about gendered qualities are certainly not shared by all 891 participants, the responses do show some consistency in their notions about what “Men are ...” and “Women are ...”. (The man-woman binary was assumed in the survey design as it is the mainstream binary gender system in the United States as well as to make results comparable to previous research).

Overall, 1091 responses to the “Men are ...” and “Women are ...” questions were recorded. To prepare the data for analysis, each reply was labeled as belonging into one of five categories. Replies not usable for further analysis were categorized as “Not analyzable” (e.g. “Women are !!”). Comments such as “Women are female humans” or “Men are male” were labeled “Synonym”. Items in both categories were excluded from further analysis. Three categories of replies were included in analysis. First, those labeled “Primed attribute”, indicating that the response contained attributes from the Likert-scale questions that participants had answered on the questionnaire right before” (e.g. “Women are friendly”, this applies to questions 5-9 from Table 28). Second, those labeled “Non-primed attribute”, i.e. the comment mentioned characteristics that were *not* part of the Likert-questions on the survey (e.g. “Women are sexy”). The label “Linguistic feature” was applied to comments picking up on linguistic features of the stimulus (e.g. “Women don’t use exclamation marks”). Note that there is no qualitative difference between the results in the categories “Primed attribute” and “Non-primed attribute”. Separate analysis was necessary only because participants offering a “Primed attribute” were potentially

influenced by exposure to the respective survey question. Counts of such primed examples (which are among the most frequent responses) could then not be productively compared to input volunteered without this kind of prompting.

To establish the topics covered in these comments that could be relevant as gender stereotypes, all content words (i.e. nouns, verbs, and adjectives) were extracted from the responses. This was achieved by tokenizing the text using the function `word_tokenize` in the Python Natural Language Toolkit and using the NLTK stop word list, a collection of English function words such as *to*, to eliminate non-content words. A Python script marking negated forms was applied to the results in order to be able to distinguish between forms such as *assertive* and *not assertive*.

This resulted in 1775 content words (out of 545 responses) for “Men are ...” and 1970 content words (out of 546 responses) for “Women are ...”. Token counts for each content word were then tallied by gender. The counts were normalized by dividing by the total number of content words left in response to the respective question. This allowed for comparison of results between the two questions. For instance, if participants left five mentions of *kind* in the “Women are ...” box, and the total number of content words extracted from answers to “Women are ...” was 100, the result would be reported as $5 / 100$, or 0.05. If we received the same number of *kinds* for “Men are ...”, but out of a total of only 10 words, the results would $5 / 10 = 0.5$. Below, results for the categories “Primed attributes” and “Non-primed attributes” are described.

5.4.3.1. Primed gendered attributes

Several participants (119 comments total) picked up on the attributes queried in earlier parts of the questionnaire. That is, they used the terms from the Likert-scale questions (e.g. *friendly*, *educated*) to describe how “Men are ...” and “Women are ...”. This

is an obvious kind of priming that had not been anticipated in the study design when those questions were placed last. The rationale behind that decision was to keep participants from guessing the purpose of the study early on.

Women are ...	Tokens (divided by total words)
<i>sensitive</i>	21 (0.01066)
<i>friendly</i>	13 (0.00660)
<i>assertive</i>	8 (0.00406)
<i>timid</i>	8 (0.00406)

Table 34: Most frequent responses to the “Women are ...” question using potentially primed attributes.

Men are ...	Tokens (divided by total words)
<i>assertive</i>	43 (0.0242)
<i>friendly</i>	8 (0.00451)
<i>insensitive</i>	5 (0.00282)

Table 35: Most frequent responses to the “Men are ...” question using potentially primed attributes.

5.4.3.2. Non-primed gendered attributes

Several participants also entered concepts not mentioned before in the questionnaire, adding to our vocabulary of potential gender stereotypes.

Women are ...	Tokens (divided by total words)
<i>friends</i>	12 (0.00609)
<i>good</i>	11 (0.00558)
<i>open</i>	11 (0.00508)
<i>kind</i>	11 (0.00558)
<i>beautiful</i>	10 (0.00508)
<i>strong</i>	8 (0.00406)
<i>social</i>	8 (0.00406)
<i>passive</i>	8 (0.00406)
<i>smart</i>	8 (0.00406)
<i>nice</i>	8 (0.00406)

Table 36: Most frequent responses to the “Women are ...” question using non-primed attributes.

Men are ...	Tokens (divided by total words)
<i>strong</i>	17 (0.00958)
<i>aggressive</i>	14 (0.00789)
<i>fun</i>	14 (0.00789)
<i>good</i>	13 (0.00732)
<i>direct</i>	9 (0.00507)
<i>great</i>	9 (0.00507)
<i>smart</i>	9 (0.00507)
<i>cool</i>	8 (0.00451)
<i>nice</i>	8 (0.00451)

Table 37: Most frequent responses to the “Men are ...” question using non-primed attributes.

To summarize, prevailing gender stereotypes in our participant pool are: Men are *assertive* (0.024), *strong* (0.00789), *fun* (0.00789), and *good* (0.00732). Women are *sensitive* (0.0106), *friendly* (0.0066), *friends* (0.00609), and *good* (0.00558). (*Friendly* and *friends* were not combined into one item to keep different parts of speech separate). These characteristics overlap to quite some extent with Broverman et al.’s (1972) male-valued “competency” cluster and the female-valued “warmth and expressiveness” cluster respectively, suggesting that their findings are still relevant to our present-day sample.

Another gender stereotype emerging out of the perception study, although it is never explicitly mentioned, is what Rich (1980) called “compulsory heterosexuality” (Butler (1999:xxix) talks about the “heterosexual matrix”): the assumption that women are romantically interested in men, and vice versa. In the perception experiment

presented here, 830 of the 891 participants perceived the author to be heterosexual. This means that for the majority of participants, the choice of gender already presupposes the addressee of a dating ad. When participants think of a female author of a dating ad, they assume a male addressee. Thus the sparsity of perception data on m4m and w4w ads. It illustrated clearly that participants are operating under a heteronormative conception of gender.

This heteronormative assumption is an important part of American gender ideologies, especially applicable to men (Cameron 1997; Kiesling 2009). Recent studies documenting this attitude in the U.S., especially among adolescents, include Nielsen et al. (2000) Tolman et al. (2003), or Renold (2006).

Based on these findings regarding stereotypical attributes and the heteronormative ideology, we will assume that the authors in the dating ad corpus hold similar gender stereotypes. While impossible to verify empirically for individual authors, this seems a reasonable assumption considering the input from the perception study and the external research cited above.

Accepting this assumption and considering the results of previous studies that engaged with language ideologies and gender (e.g. Hall 1995; Eckert 1996; Kiesling 2009), we expect to see the linguistic performance of authors to index these gendered attributes (*assertive, strong, etc.; friendly, sensitive, etc.*) as part of constructing a gendered persona (Eckert & Rickford 2001). In dating ads, authors can orient themselves towards such stereotypically gendered characteristics when presenting themselves as a man, a woman, or any other gender category (where “orienting towards” includes the option of rejecting them). Their readers, on the other hand, will read the ads through the lense of gender stereotypes and the heteronormative assumption. The question emerging out of this discussion of gender stereotypes is thus: How do authors engage with these

stereotypes when creating a gendered identity for a specific audience in their dating ad, constrained by genre conventions?

5.4.3. Social meaning of emoticons and prosody

This discussion of the social and linguistic context allows us to return to our discussion of the indexicalities of the two gendered features in our dataset. For each, it will highlight the interactions between results of the perception study and the production study in light of the context effects outlined above.

5.4.3.1. Emoticons

First, the social meaning of emoticons will be considered. The label *emoticon* itself suggests two relevant characteristics of these items: they are about emotions; and they are icons, in the sense of Peirce (1940), who explicitly distinguishes the icon from the index. Emoticons, essentially pictographs of the human face, are thus qualitatively different from the linguistic variants commonly discussed in the literature on indexicality such as Eckert (2008a). However, in the present study, emoticons have apparently acquired an indexical meaning to the study participants: they index female gender.

Following the approach by Ochs (1992) outlined above, we establish their indexicality as follows, focusing on the most frequently used emoticon in the present dataset, the :). As will be shown below, :) is representative of the entire feature set. Like all emoticons, :) expresses the author's emotional state – in this specific case, a positive emotion. Perception studies show that its indexical field includes characteristics such as “happy”, “honest”, or “surprised” (Walther & D’Addario 2001). The perception study indicates a weak correlation with “friendly”.

Some of these readings seem consistent for the way the emoticon is used in the dating ads corpus: it occurs most frequently during introductions by both w4m authors

(*Hey there good looking. :) <corpus file 1030636>*) and m4w authors (*Hi there :) <corpus file 1082695>*) or to mitigate requests: for example *I love men in uniform :) (<corpus file 1039986>*) from the w4m category, or *But be somewhat attractive :) (<corpus file 1081626>*) from the m4w category. While these might be preferred contexts, this feature really can occur in most post-sentence positions (for instance: *But I do thank you for reading :) <corpus file 1082652>*) or: *I'm five eleven 195 farmer shoulders :) <corpus file 1028616>*). All instances inspected are consistent with the “happy” or “friendly” meaning. The second most frequent item, ;) is used pretty much interchangeably with :) except that it does not occur in introductions. This is somewhat unexpected given the results in Walther & D'Addario (2001) which show a different perceptual value along the lines of “secretive”, “sarcastic”, and “seductive”.

The “happy” and “friendly” indexical meaning of emoticons ties neatly into the system of gender stereotypes discussed above: Broverman et al.'s (1972) results indicate that stereotypical men “always hide their emotions” while women “never hide their emotions”. The :) emoticon matches the characteristics in the “warmth and expressiveness” cluster of characteristics that according to Broverman et al. (1972) are perceived as both typical of and desirable in women. In the perception study, participants similarly indicate that “friendly” and “sensitive” are the most strongly female-associated stereotypical attributes. Similar readings apply to almost all emoticons found in the dating ads corpus, as they all express positive emotional affect (a list of full results is presented in section 3.3.1.4.). On the other hand, emoticons do not speak to the characteristics “strong” and “assertive” that emerge as most strongly male-gendered among the gender stereotypes.

Thus, in the present dataset emoticons index characteristics such as friendliness and emotional expressiveness stereotypically associated with women.

5.4.3.2. Prosody

A very similar case to the one presented for emoticon use can be made for the prosodic items, the most frequent of which are variations on *haha* (this includes *hahaha*, *hah*, etc.) and *ya* (including *yayay*, *yah*, etc.). This feature too signals positive emotional expressiveness (one could argue that *haha* is a character-based version of the smiley). In the dating ads corpus, *haha* is used to represent laughter, either to signal a joke (*and job like be a doctor or gold miner haha* <corpus file 1039856>) or to convey positive emotion (*haha Steve Harvey is Hah-Larious!* <corpus file 1027613>) and thus similar to the emoticon use above (indeed, the two can be combined: *Message me :) haha* <corpus file 102291>). Just like with the emoticons, this indexical connection fits into Broverman et al.'s (1972) "warmth and expressiveness" cluster. It fits the gender stereotypes "friendly and "sensitive" established in the perception study. Again just like emoticons, prosodic items are weakly positively correlated with friendliness in the perception study. It is noteworthy that several participants in the perception study in their comments describes this feature (*haha*, and *sooo* in the stimulus), as "flirty".

Unlike the emoticon stimulus, results from participants exposed to the prosody stimulus show a significant difference for the perceived sensitivity by perceived gender: female author-perceived ads were rated as significantly more sensitive than the male-perceived ones (mean when author gender = male: 3.17, when author gender = female: 2.77; $t = -2.48$, $p = 0.017$). That is, participants gave a lower score (where 1 = very sensitive) because they thought the author was a woman, rather than because they felt the prosodic items to index sensitivity no matter the author gender. This might, however, be an effect of perceptual salience, an issue discussed in section 5.7.2. below.

Summing up the results on the two gendered features, we find that both can index stereotypically female-gendered attributes such as friendliness and emotiveness.

Matching their perceptual value and their use in linguistic production to gender stereotypes present in the population allowed us to account for the specific indexical meanings activated in this context. We can test this indexical relationship empirically within the participants in our dataset: in the mind of participants, are emoticons and prosodic items more strongly correlated with a specific gender or with a specific attribute, such as friendly? For instance, when confronted with the emoticon stimulus, do participants rate the text author higher on friendliness because they perceive the emoticon as indexing friendliness, or do they rate the author higher on friendliness because the emoticon tells them the author is female and therefore friendly? In short: does the text stimulus or the perceived gender better predict scores on attributes like friendliness? Over the entire dataset, participants who perceived the author to be female tended to perceive the author as slightly more friendly, a difference that is not statistically significant (mean across all stimuli is 1.83 for female-perceived, 2.03 for male-perceived authors; note that 1 = very friendly, 5 = very unfriendly). Female-perceived authors are perceived as slightly more sensitive except for one stimulus: the mean score on sensitivity, where 1 = very sensitive, is 2.74 when the author was perceived to be female and 2.85 when the author was perceived to be male, another non-significant difference.

Regression models for each attribute with perceived author gender and stimulus type as predictors suggest that the stimulus is a non-significantly better predictor for assertiveness and education – that is, these features vary slightly more by the stimulus rated by the participant, independent of perceived gender. Regression models were compared using the Akaike information criterion (AIC), a statistic that estimates how much information is lost when the data is modeled in a specific way (Crawley 2007:353). In this way, it can be used for model comparison, where a lower AIC score indicates a better model. For perceived assertiveness, the gender-predicted model (AIC = 2238.1)

was better than the stimulus-predicted model (AIC = 2237.8). Similarly the gender-predicted model (AIC = 1787.5) was better than the stimulus-predicted model (AIC = 1772.7) in modeling results for perceived education. Friendliness and sensitivity perception, on the other hand, are slightly better predicted by perceived author gender, but again, there is not statistically significant difference. In these two cases, the fact that the author was perceived to be female was more predictive of a low friendliness and sensitivity score (where 1 = very friendly, very sensitive) than the linguistic feature the participant was exposed to (perceived friendliness, gender-predicted: AIC = 1776.5, stimulus-predicted: AIC = 1799.2; perceived sensitivity, gender-predicted: AIC = 1727.4, stimulus-predicted = 1743.9).

The fact that none of these differences are statistically significant or approaching any kind of practical significance supports the shifting and fluid nature of indexicals as postulated for instance by Ochs (1992). It suggests that indeed the two indexicalities are intertwined to an extent that makes it hard to empirically identify the more important one.

5.5. INDEXICALITY OF NON-GENDERED FEATURES

None of the other e-grammar features investigated here hold strongly gendered meanings to the participants in the perception study. In the production study, they do not pattern similarly to the gendered emoticons and prosodic features either. Nor is their perception regarding social attributes other than gender quite conclusive, except that all of them, except for repeated punctuation, correlate weakly positively, but not significantly, with friendliness and weakly negatively with education.

5.5.1. Capitalization and repeated punctuation

As discussed before, however, there is a clear functional divide among the remaining e-grammar features. All but two of them are abbreviation devices that can be understood as attempts to save space and time. Those two exceptions are capitalized words and non-Standard punctuation. In both cases, writers invest extra effort to include them in their ad; presumably, to make a point. But what meaning are they conveying? Participants in the perception experiment did not associate them with a specific author gender or attribute. (Actually, they are the feature stimuli closest to the control stimulus).

In the production study, however, capitalized words and repeated punctuation pattern pretty much identical, with m4m and w4w authors grouped closely together, leading in feature use; the w4m and m4w authors are below the mean frequency for both features. This ordering does not make sense under the audience design model as applied to emoticons and prosodic items above; the close similarity between m4m and w4w authors contradicts the model. Additionally, since there is no clear outcome of the perception study regarding their social meaning, we cannot confirm them to be the kind of gender marker speakers can accommodate towards.

A look at the details of the results regarding capitalization suggests that looking at frequencies only is misleading: in this case, w4w and m4m writers use capitalization to the same extent but in different ways. When looking at the most frequently used tokens within each category, we see that capitalized forms fulfill very different functions within these groups.

Capitalized item	Tokens
<i>DDF</i>	1,315
<i>YOU</i>	700
<i>NOT</i>	662
<i>HIV</i>	654
<i>AND</i>	604
<i>HWP</i>	567
<i>NSA</i>	485

Table 38: Most frequently capitalized words in ads from the m4m category.

Capitalized item	Tokens
<i>MEN</i>	1,437
<i>NOT</i>	690
<i>BBW</i>	583
<i>COUPLES</i>	567
<i>PIC</i>	348
<i>PLEASE</i>	323
<i>YOU</i>	302

Table 39: Most frequently capitalized words in ads from the w4w category.

The items listed suggest that these two groups do use capitalizations frequently, but that they do so for different purposes and in different forms. For the m4m authors,

the most frequently capitalized item is the abbreviation *DDF* (drug and disease free), for w4w it is *MEN* – usually used in sentences along the lines of *NO MEN PLEASE*. (Note that the most frequent items in both categories outnumber the second-ranked item by an order of magnitude.) The results suggest that w4w authors tend to capitalize entire words to emphasize their message (use capitalization to shout), with other frequent words including NOT and COUPLES. The m4m authors, on the other hand, use capitals mainly in abbreviations (use capitalization to abbreviate). The m4m counts for capitalized words are thus correlated with their frequent use of abbreviations, where they lead all other groups in feature frequency. In hindsight, excluding the abbreviations from the capitalization results in the production study might have been instructive. But, as pointed out before, authors have a choice to use lowercase letters for abbreviations just like with any other word.

This analysis of usage patterns of capitalized words suggests that capitalizing words does not have a shared meaning across those groups that could map to gender in any way. Capitalization is frequent among w4w and m4m authors, but for very different reasons: in one case, they are most frequently used to abbreviate words, in the other case mostly for emphasis.

Regarding punctuation, the picture is not as simple. Both m4m and w4w mainly use tokens of <...>. To a lesser extent, writers from both groups also use multiple exclamation marks.

Punctuation type	Instances
.	16,009
!	2,349
+	1217

Table 40: Most frequent non-Standard punctuation patterns in the m4m category.

Punctuation type	Instances
.	11,674
!	2,904
*	345

Table 41: Most frequent non-Standard punctuation patterns in the w4w category.

Usage patterns of these punctuation items do not differ very much between the two categories m4m and w4w. Multiple punctuation is used to end a sentence, effectively replacing a single <.> (e.g. *I dont smoke, I do drink socially ..* <corpus file 1001095>) or, less frequently, within a sentence (e.g. *I'm loyal .. like really loyal* <corpus file 1001201>). Some frequent use cases are to use multiple stops to end an ad (*hit me up ..*, <corpus file 1086853>), or conclude the initial greeting (*Hello ..* <corpus file1087511>), or to end a list of things (*dungeons and dragons, etc ...* <corpus file 1087663>). The number of stops does not seem to make a difference in meaning or use. The overall effect of this device seems to be to mirror pauses in speech (consider *I'm loyal .. like really loyal* versus *I'm loyal*

like really loyal versus *I'm loyal. Like really loyal.*) much in the way the standard full stop does (Chafe 1988). Similarly, writers use multiple exclamation marks as more emphatic variants of the single exclamation point (*I'm not changing for anyone!!* <corpus file 1013669>). In both cases, the Standard meaning of the punctuation item is amplified by repetition. Looking at the two third-ranked items, it is noticeable that, in contrast to full stops and exclamation marks, they are not punctuation characters in Standard written English. The asterisk *, third-most frequent in the w4w category, is used much like an exclamation mark to emphasize a point. Again, this is often related to the “no men” statements discussed above (****females only**** <corpus file 10007724>), less frequently to censor content (*I can be a B***** <corpus file 108674>). Combinations of +, third-most frequent type in the m4m category, are generally used in m4m ads to indicate approval or desirability in the sense of “very good” (*45 or over +++++* <corpus file 1088453>) or to list items (*Frat +++ White +++ Muscle +++* <corpus file 1089080>). Note, however, that both * and + are niche cases and are substantially less frequent than stops or exclamation marks.

In summary, no difference in use or relevance of non-Standard punctuation between the m4m and w4w groups is apparent. The similarity in patterning between capitalizations and repeated punctuation, on the other hand, is a coincidence rather than an indication of linguistic similarity.

5.5.2. The gendering of Standard forms

Returning to the remaining e-grammar features, namely abbreviations, clippings, rebus forms and single letters, we may note that in addition to being character- and timesaving devices (as discussed earlier), they are all non-Standard forms of writing. They violate the norms of written English as used in the media and codified in dictionaries. We

might then analyze them under this broader heading: is there a logic to how writers use and readers perceive non-Standard features? Since all of the e-grammar features under investigation here can be considered non-Standard, the following section will address this question in more general terms and include all of the features discussed above.

To restate: a unifying aspect of all e-grammar features is the fact that they are not part of Standard written English. This is something that is frequently commented on in the dating ads corpus, where we find comments such as [*Looking for someone who c]an spell words without using "text talk" style.* (<corpus file 103143>) or **If you can't spell correctly, please do not hit me up** (<corpus file 102794>). Several language-external factors conspire to make Standardness a thorny issue in online written language.

First, written language is more explicitly and thoroughly standardized than speech (Milroy 2002:47); indeed, for many speakers, the written word is the definition of a Standard form. Norms of the written Standard are codified in dictionaries, taught in schools, and enforced by automatic spell checkers. On the other hand, the more informal, speech-like quality of computer-mediated communication writing has been discussed at quite some length in the sociolinguistic literature (e.g. Herring 1999; Tagliamonte 2012). It is thus hard to define what should be considered Standard or non-Standard within these ads. (Incidentally, the distinction is often equally blurry in spoken language). However, most English-speakers do not consider deviation from the codified Standard acceptable in any setting (Milroy 2002) and comments from ad authors like the ones cited above indicate that this attitude is present among at least some writers in the dating ads corpus as well. Another indicator that e-grammar features are perceived as non-Standard forms is found in the perception study results. Note that the control stimulus, designed to be as “feature- and gender neutral” as possible, is also the most Standard: test stimuli were created by adding e-grammar, that is non-Standard, features. The results of the

perception study then can also be interpreted as more general reactions to adding non-Standard linguistic forms to a piece of Standard writing. From this perspective, the fact that every treatment stimulus is lower, though non-significantly so except for the clippings and the prosody stimulus, than the control stimulus in perceived education could be instructive: these low education scores correlate weakly with high scores on “friendliness”. This is a pattern consistently found in matched guise studies such as Campbell-Kibler (2008): speakers are perceived as less educated but friendlier when they use non-Standard features. It is thus possible that the same dynamic is reflected in the results of the perception study. In summary, it is possible that participants in the perception study react to the independent latent variable “Standardness” in addition to picking up on gender or other social characteristics of the author.

If we assume this pervasive Standardness effect for the perception study, we can relate it to another aspect of the data that has not been discussed so far. Two items were identified as female-indexing; Standardness might offer an approach to exploring *male*-gendered features.

In the perception study, all e-grammar, non-Standard stimuli are perceived as more likely to be female-authored than the control stimulus, albeit only two of them significantly so. The control stimulus itself attracts significantly more “the author is male” ratings than “the author is female” ratings.

If non-Standard features are consistently perceived as more feminine (note, however, that the difference is statistically significant for prosody and emoticons only), can we make the case that Standard writing style indexes male gender? The discussion below suggests two paths to analysis and potential further research avenues, rather than an explanation; the data at hand does not suffice for a clear, empirically supported finding.

First, e-grammar features might be perceived as innovative. These e-grammar features, some of them labeled “text speak” elsewhere, are commonly perceived as new and incoming variants (Tagliamonte 2016). In the sociolinguistic literature, female speakers have quite consistently been found to use more innovative features (Labov 1990). It could then be argued that this fact motivates participants to perceive anyone using features perceived as incoming to be female. However, contrary to these empirical findings (which also pertain to non-proscribed variants only, see the discussion of the “gender paradox in 2.3.1.), the public perception is apparently that women are more conservative and “correct” in their language use (Trudgill 1972). In this case, we would expect the outcomes of the perception study to be exactly opposite of what they are.

Alternatively, it is interesting to note that the author of the control stimulus, in addition to being perceived as more educated and less feminine than the treatment stimuli, is also rated comparatively high on assertiveness, even though no statistically significant differences can be found. Assertiveness was the most frequently mentioned male gender stereotype in our study and is considered a male trait in Broverman et al (1972), strengthening the case that the Standard stimulus is associated with traditionally male attributes.

Considering the results of the production study in conjunction with the perception study, it must be noted that the m4w authors score below the mean for all non-Standard features (except in the female-gendered prosodic items and emoticons, where we argued for an accommodation effect above). They indeed seem to be very Standard in their language use when it comes to e-grammar features. The m4m authors, on the other hand, consistently score above the mean: they use a lot of non-Standard features. This conflicts with the idea that Standard features are somehow indicative of male writers in general. However, we might hypothesize that Standard features index attributes deemed desirable

in male writers, analog to friendliness and emotional expression for the emoticons and prosody items.

Standard English forms are, according to the language perception research cited above and to some extent the perception study, perceived as indicative of educational attainment. The fact that m4w writers would want to include this indexicality in their gender performance, and m4m writers would not, could be explained by referencing the social exchange theory discussed in section 3.1.1.. Social exchange theory posits that dating ads set up a social exchange. In the context of heterosexual dating, this exchange – simply put – consists of men offering social status and professional success in exchange for women’s physical attractiveness and youth. In this scenario, it would make sense for men to index educational attainment, an attribute closely linked to professional success, when trying to set up this kind of exchange. The m4m authors, presumably, do not participate in the same kind of exchange. This approach to analyzing potentially male-gendered features does not constitute conclusive analysis; rather, it shows yet another way of how consideration of social context could affect the use of language forms and point to possible paths of future research on gendered language variation.

5.6. PERCEPTUAL ATTRACTIVENESS OF E-GRAMMAR FEATURES

To assess the effectiveness of using gendered features in presenting a desirable female or male identity, a question asking participants to rate the perceived attractiveness of the author was included in the perception survey. This was implemented as a Likert-scale question ranging from “Very attractive” to “Very unattractive”. The hypothesis to be tested was that text stimuli containing gendered features would be rated as more attractive by the target population (i.e. men for female-gendered features, women for male-gendered features). For instance, a stimulus containing emoticons or prosodic items

would be perceived as more attractive than the control stimulus by participants who self-identified as male and heterosexual. However, no significant differences could be established for any of the features under investigation, neither for emoticons, prosody, nor any other features.

5.7. METHOD EVALUATION

Two issues need to be kept in mind when evaluating the results of this analysis. Those concern the theoretical status of the e-grammar features studied here and the question of salience in the perception study.

5.7.1. E-grammar features in sociolinguistic theory

Variationist sociolinguistics, following Labov (1978:7), traditionally defines the sociolinguistic variable as “two ways of saying one thing”. This kind of variable is the central object of study in classic sociolinguistic research such as Labov (1966) or Trudgill (1974). Since most of the previous research cited here as well as the present study is situated in this research tradition to some extent, a short note on the status of our features in this regard is appropriate. (Note, however, that the idea of two variants “meaning the same thing” has already been weakened in third-wave sociolinguistics studies (Eckert 2012a): the research agenda’s whole point is that the social meaning of variants can differ even if their dictionary definition is identical). The idea of the sociolinguistic variable has also changed along with the field’s methodological shift towards conceptualizing variation as gradual rather than discrete. That is, a variable need not be categorized as one of “two ways” but can fall somewhere on a range of measured values. Sociolinguistic research in general focused on sounds; what exactly a variable looks like in writing has not been as well defined (Lillis 2013). Intuitively, differences in spelling or formatting of a word might be considered equivalent to the differences in sound production typically

studied in the sociolinguistic literature. In our feature list, this would apply to capitalized words (with a (word)'s variants <word> and <WORD>), clippings ((information): <information>, <info>), leetspeak ((the): <the> and <th3>), rebus forms ((for): <for>, <4>), and single letters ((you): <you>, <u>). The status of abbreviations such as *LTR*, short for *long term relationship*, seems to fall into the same category. However, this does not seem to be the case for some shortenings such as *LOL*. While this technically stands for *laughing out loud*, the two forms are not used interchangeably since the long form is very rare and *LOL* has essentially been lexicalized as its own word. (And is recognized as such in the *Oxford English Dictionary*: “used to draw attention to a joke [...], or to express amusement”.) The punctuation feature, which mostly consists of re-duplication of <.>, <!> and <?> is another borderline case. Does <...> mean the same as <.>? And how about <????> and <?>? Emoticons and prosodic features prove similarly hard to categorize: they are typographic representations of facial expressions and non-linguistics sounds, aspects of interaction that Labovian sociolinguistics has paid little attention to. The two features do, on the other hand, meet Labov’s standard in that they do not change the semantic meaning of a sentence; they could be operationalized as binary (present-not present) variants – much like, for instance, non-prevocalic (r) in Labov’s studies in New York City.

Also note that in this study features are not has quantified as a percentage of variant A versus variant B, as in the traditional phonological variation study. Rather, feature frequency is measured as token count normalized by word count, a measure commonly employed in corpus linguistics (Biber, Conrad & Reppen 1998) or computational linguistics (Bamman, Eisenstein & Schnoebelen 2014). This is due to the size of the dataset and the nature of some e-grammar features: for instance, how would one conclusively determine where the author could have used an emoticon, but did not?

Maybe even more importantly, the study here focuses on bundles of features, each comprising intuitively similar features rather than one well-defined variable. While care was taken to inspect and understand individual items in context, this is another reason the feature set cannot be considered quite equivalent. However, the fact that this study addresses a new set of features in a less well-researched language mode made this an appropriate approach to analysis.

These methodological differences distinguish the approach taken here from the Labovian paradigm.

The e-grammar features, on the other hand, fit in neatly with the language and gender research paradigm (e.g. Holmes 1990; Coates 1991) that has mostly focused on the presence of absence of features such as question tags or other lexical items.

5.7.2. Perceptual salience

Another issue that has been brought up by perception studies working with text (Queen & Boland 2015) is the issue of salience. Which features do participants pick up on, which are not noticeable? For the present dataset, we can say with certainty that at least some participants pick up on the emoticons and the prosodic items. This is obvious from the comments left on the survey and the fact that they pattern differently from the control stimulus in a consistent fashion. These two most strongly gender-linked features in the dataset are also the intuitively most salient ones. For the other features, we cannot be sure if participants do not notice them, for instance because they appear in isolation rather than part of a consistent style, or just do not attach any meaning to them. It must be kept in mind that linguistic features, outside of this experimental setting, usually co-occur with other items as part of a linguistic style, rather than in isolation (see also the discussion of the stimulus creation in chapter 3). It is very well possible that participants would have,

to name two random features, perceived a stimulus containing capitalized words and non-Standard punctuation as exhibiting a distinctively male style. The study design and limited resources did not allow for adding that amount of complexity, but the point must be kept in mind.

It also needs to be noted that some of the features that are not significantly different from the control stimulus – that is, all features besides the emoticons and prosodic items – might very well be meaningful to at least some participants. A closer analysis of the impact of participant demographics than possible during this project might help clarify this issue.

5.8. CONCLUSION AND IMPLICATIONS

In this chapter, the results of a linguistic production and perception study were used to present an account of gendered variation in computer-mediated communication as well as the indexical value of emoticons and prosodic items emerging as gendered in this context. To this end text genre, audience effects, and gender ideologies were discussed and brought to bear on the results of the quantitative analysis.

Two features are identified as gendered in the context of online dating ads: emoticons, such as :), and prosodic items, such as *haha*. Both are on average produced more frequently by female writers and are also, as the perception study indicates, perceptually linked to female author gender. None of the other e-grammar features – which include the use of abbreviations, clippings, capitalized words, non-Standard punctuation, and the substitution of words with single letters or numbers – show a similarly gendered stratification. The indexicality of the two gendered features is established by a thorough study of their linguistic and social context, which suggests that they tie into American gender stereotypes, which include the belief that women are

friendlier and more emotionally expressive than men. It is shown that the use of these features is also highly dependent on addressee gender.

Regarding the study of language and gender, the study has pointed to the relevance of addressee effects, more specifically addressee gender effects, which need to be taken into account in language and gender research. It also suggested ways of conceptualizing gender other than the traditional male–female binary, while still working with a sizeable dataset.

The study showcases several computational techniques, such as k-means cluster analysis for pattern identification or a machine learning algorithm used to establish word groups, and ways in which they can productively be used in sociolinguistics. This computer-assisted approach to sociolinguistic research might prove especially valuable for the analysis of large text corpora increasingly used in the study of language in computer-mediated communication.

As the study has shown, future work studying the social meaning of linguistic forms ought to remain cognizant of the interplay of linguistic production, perception, and social context such as genre conventions or social ideologies. The study sketches various approaches to studying each of these constructs empirically, and how these analyses can then be brought to bear upon each other. The results emphasize that the indexicality of a linguistic feature cannot be understood without paying close attention to the social environment out of which its meaning emerges. This suggests that assigning a gendered or other social meaning to a linguistic variant without considering the interaction of linguistic production, linguistic perception, and social ideologies surrounding it might be problematic. A minimal list of the criteria to be met includes the following:

- A dataset that foregrounds the social aspects under investigation must be chosen.

- The relevant social groups and linguistic features cannot be presupposed, but should be established through empirical analysis.
- In order to assign social meaning to a variant, analysis of perception and production must exhibit consistent results: the feature must exhibit social relevance in both settings.
- Feature set and aspects of context such as genre and locale should be as consistent as possible across studies of production and perception.
- Results should be statistically significant to be considered reliable. Patterns established through visual inspection can lend further support, but should not be considered conclusive.
- A thorough analysis of social and linguistic context needs to inform the argument about which indexical meaning is activated in the given context. This analysis should rely on empirical findings, preferably combining external research with analysis of the dataset at hand.

Appendix

A.1. CLUSTERTOOLS SAMPLE OUTPUT

Last login: Fri Jul 8 21:09:15 on ttys000

dhcp-128-83-211-47:~ ps22344\$

*/var/folders/n_/d_v1xh692r130s5_mr35r9zrn9nbc7/T/Cleanup\ At\
Startup/02_analysis_cluster_0501-489857404.291.py.command ; exit;*

[...]

[...]

Input statistics

std

[0.0017387 0.00154774 [...] 0.00076621]

min

[0. 0 0. 0. 0. [...]]

max

[0.32132565 0.01858736 0.01123596 [...]]

means

[3.59702603e-05 6.04842085e-04 6.82398247e-05 [...]]

median

[0. 0[...]]

range

[0.32132565 0.01858736 0.01123596 [...]]

[...]

Working with manhattan distance metric

CLUSTERING CALLED *AgglomerativeClustering*([...]) HAS 4 CLUSTERS
Its silhouette score is -0.0129067210144

Cluster 0 contains 10356 items, 27.0 % of the total
2376.0 items of category w4w make up 23.0 % of this cluster
2062.0 items of category w4m make up 20.0 % of this cluster
2426.0 items of category m4w make up 23.0 % of this cluster
3492.0 items of category m4m make up 34.0 % of this cluster

Cluster 1 contains 6280 items, 16.0 % of the total
1949.0 items of category w4w make up 31.0 % of this cluster
1888.0 items of category w4m make up 30.0 % of this cluster
1779.0 items of category m4w make up 28.0 % of this cluster
664.0 items of category m4m make up 11.0 % of this cluster

Cluster 2 contains 14458 items, 38.0 % of the total
3037.0 items of category w4w make up 21.0 % of this cluster
3001.0 items of category w4m make up 21.0 % of this cluster
3401.0 items of category m4w make up 24.0 % of this cluster
5019.0 items of category m4m make up 35.0 % of this cluster

Cluster 3 contains 7015 items, 18.0 % of the total
2108.0 items of category w4w make up 30.0 % of this cluster
2477.0 items of category w4m make up 35.0 % of this cluster
1774.0 items of category m4w make up 25.0 % of this cluster
656.0 items of category m4m make up 9.0 % of this cluster

Statistics per category

Category w4w has 9470 items
2376 items or 25.0 percent in cluster 0
1949 items or 21.0 percent in cluster 1
3037 items or 32.0 percent in cluster 2
2108 items or 22.0 percent in cluster 3

Category w4m has 9428 items

2062 items or 22.0 percent in cluster 0
1888 items or 20.0 percent in cluster 1
3001 items or 32.0 percent in cluster 2
2477 items or 26.0 percent in cluster 3

Category *m4w* has 9380 items
2426 items or 26.0 percent in cluster 0
1779 items or 19.0 percent in cluster 1
3401 items or 36.0 percent in cluster 2
1774 items or 19.0 percent in cluster 3

Category *m4m* has 9831 items
3492 items or 36.0 percent in cluster 0
664 items or 7.0 percent in cluster 1
5019 items or 51.0 percent in cluster 2
656 items or 7.0 percent in cluster 3

Strongly predictive features are

Raw Scores

Cluster 0 and cluster 1 are differentiated by

i : -0.00909424943968, ... : 0.00199315043929, *you* : -0.00132574709377, *me* : -
0.000834264067738, *text* : 0.000721702663683, *someone* : -0.000672850717223, *it* : -
0.000664128982281, *like* : -0.000598446540347, *know* : -0.000579941023535, *who* : -
0.000561192932894

Zscores

Cluster 0 and cluster 1 are differentiated by

i : -4.9083144259, ... : 4.19929633224, *text* : 3.33045325461, *looking* : 2.57160624027, *age* :
2.00085325436, *you* : 1.07379132631, *pic* : 0.913935139898, *im* : 0.764246463974, *send* :
0.753506701896, *we* : 0.68758774653

[...]

Here is a typical document for each cluster

We set the distance metric to cityblock

CLUSTER 0

```
<file> <title=submissive here looking to obey> <plat=ad> <city=baltimore> <date=2015-06-18> <time=9:49am>
[...]> <text> I like to be used for your pleasure.anything you want [...] </text> </file>
```

[...]

Comparing clusterings

CLUSTERING CALLED <class 'sklearn.cluster.hierarchical.AgglomerativeClustering'>
HAS 4 CLUSTERS

Its silhouette score is -0.0129067210144

CLUSTERING CALLED <class 'sklearn.cluster.hierarchical.AgglomerativeClustering'>
HAS 8 CLUSTERS

Its silhouette score is -0.0270833882259

[...]

Metric: adjustedrand_sim

	<i>AgglomerativeC lustering--16</i>	<i>AgglomerativeC lustering--24</i>	<i>AgglomerativeC lustering--4</i>	<i>AgglomerativeC lustering--12</i>
<i>AgglomerativeC lustering--24</i>	0.948301368	***	0.70435312	0.9351187
<i>AgglomerativeC lustering--4</i>	0.75227497	0.70436212	***	0.76470697
<i>AgglomerativeC lustering--12</i>	0.9867677	0.935113187	0.76497	***
<i>AgglomerativeC lustering--8</i>	0.88866049	0.8378942	0.859748438	0.9018489

[...]

A.2. WORD2VEC SEMANTIC GROUPS

0 : POSITIVE PERSONAL CHARACTERISTICS

[u'absolute', u'beauty', u'financial', u'wisdom', u'ambitions', u'substance', u'opinions', u'unique', u'individuals', u'monogamy', u'support', u'spiritual', u'believer', u'believes', u'values', u'aspects', u'chivalry', u'condition', u'accept', u'ethics', u'belief', u'spiritually', u'appreciation', u'self', u'accepting', u'understands', u'admire', u'seek', u'maturity', u'opinion', u'confidence', u'devotion', u'emotionally', u'integrity', u'compassion', u'appreciate', u'interests', u'strength', u'emotional', u'physical', u'appeal', u'supporting', u'heart', u'intellectually', u'compatibility', u'appreciates', u'importantly', u'truly', u'physically', u'cares', u'mentally', u'outlook', u'beliefs', u'emotions', u'strong', u'respect', u'flaws', u'takes', u'manners', u'utmost', u'honor', u'faith', u'himself', u'depth', u'faithfulness', u'herself', u'based', u'sensuality', u'importance', u'goals', u'moral', u'loyalty', u'maintain', u'personal', u'balance', u'qualities', u'human', u'character', u'dignity', u'patience', u'aware', u'fearing', u'soul', u'stability', u'possess', u'honesty', u'boundaries', u'capable', u'aspect', u'important', u'god', u'characteristics', u'responsibility', u'respect', u'compromise', u'believe', u'openness', u'themselves', u'kindness', u'attributes', u'appearance', u'value', u'supportive', u'worthy', u'esteem', u'sarcasm', u'centered', u'dedication', u'trust', u'intellect', u'level', u'morals', u'speaks', u'realistic', u'communicate', u'communication', u'mental', u'lack', u'intelligence', u'higher', u'essential', u'traits', u'issues', u'ambition', u'sincerity', u'goal', u'ability']

1 : NOT CONSISTENT

[u'tame', u'foul', u'marriedattached', u'stereotypical', u'tying', u'hoe', u'ho', u'wing',
u'origin', u'fir', u'capricorn', u'smh', u'enormous', u'atm', u'absorbed', u'agressive', u'sexist',
u'lmao', u'mamas', u'pressures', u'sista', u'rolled', u'adventurer', u'rednecks', u'diff', u'frm',
u'milfs', u'doorpat', u'supper', u'hah', u'hav', u'crowd', u'creed', u'umm', u'anrabf', u'dental',
u'raped', u'jane', u'transgender', u'models', u'taurus', u'buddys', u'macho', u'recipe',
u'educate', u'freaks', u'bigay', u'locks', u'dyke', u'haveing', u'filthy', u'wheelchair', u'sa', u'sd',
u'pushover', u'bashful', u'records', u'matched', u'troll', u'dunno', u'butterfly', u'ampamp',
u'remarks', u'pisces', u'horney', u'doms', u'millionaire', u'grooming', u'anti', u'twig',
u'consultant', u'girth', u'bros', u'nothin', u'jst', u'starving', u'biased', u'flawed', u'legged',
u'technically', u'beginners', u'rage', u'abuser', u'ah', u'novice', u'vocabulary', u'herbs',
u'tricks', u'groom', u'beware', u'ts', u'ti', u'anywho', u'gosh', u'select', u'abd', u'intended',
u'severely', u'hermit', u'gfs', u'darling', u'ftm', u'ness', u'major', u'mmmm', u'heterosexual',
u'exp', u'ugh', u'gold', u'marrage', u'sayin', u'closely', u'nope', u'pagan', u'todaytonight',
u'bitchy', u'funn', u'leaning', u'triple', u'chase', u'sounding', u'cigar', u'stack', u'downlow',
u'rubenesque', u'swinger', u'aloha', u'bt', u'primarily', u'ut', u'unsafe', u'\u2661', u'fighter',
u'hella', u'crossdressing', u'softer', u'freakin', u'obesity', u'rack', u'leo', u'les', u'lez', u'maker',
u'mommas', u'aspiring', u'pigs', u'ties', u'thieves', u'princesses', u'beginner', u'bums',
u'meth', u'greedy', u'grandpa', u'painter', u'onetime', u'restrictions', u'tlc', u'newbies',
u'figures', u'ck', u'incest', u'hottest', u'discretely', u'pretentious', u'insanely', u'pitt',
u'excessively', u'tomboyish', u'matthew', u'polar', u'thief', u'nonsmokers', u'suited',
u'pansexual', u'strbi', u'bearing', u'nurses', u'hpv', u'badass', u'ciao', u'undercover',

u'bipolar', u'lonesome', u'aka', u'sarah', u'tryin', u'ir', u'muffin', u'slimmer', u'hispanics',
u'ther', u'barrier', u'minus', u'anatomy', u'bracket', u'keywords', u'dp', u'covering', u'wo',
u'handful', u'cop', u'brat', u'stink', u'bunny', u'nit', u'mamma', u'photographer', u'boxer',
u'hoping', u'poss', u'wan', u'winning', u'lactating', u'vegan', u'kitten', u'handyman',
u'suckers', u'douche', u'houseboy', u'noo', u'adrenaline', u'eh', u'challenged', u'ed', u'et',
u'pills', u'er', u'furthermore', u'xo', u'flipping', u'xd', u'silent', u'mostly', u'interacting',
u'daynight', u'raunchy', u'hoes', u'seekers', u'hve', u'gemini', u'agnostic', u'dancers',
u'dadson', u'spectacular', u'offence', u'curvier', u'babysitter', u'wolf', u'sry', u'hire',
u'carpenter', u'lonly', u'messy', u'newbie', u'referring', u'learner', u'cleandisease', u'perry',
u'ff', u'fk', u'fo', u'bcuz', u'unhealthy', u'thickcurvy', u'wacky', u'pun', u'pup', u'warts',
u'jordan', u'dtf', u'promiscuous', u'kat', u'wats', u'dealbreaker', u'cowgirl', u'stalker',
u'meanwhile', u'locals', u'experimental', u'handjobs', u'havin', u'ruff', u'housekeeper',
u'donot', u'mmmmm', u'eater', u'namaste', u'ordered', u'amounts', u'creeper', u'gq', u'gd',
u'psycho', u'felon', u'booth', u'decreet', u'vampire', u'conceited', u'fare', u'assplay', u'trashy',
u'childless', u'bee', u'coke', u'fisted', u'thatd', u'soldiers', u'winner', u'libra', u'hassles', u'tbh',
u'cuts', u'unshaven', u'thang', u'fireman', u'feelin', u'femms', u'mistaken', u'duh', u'cleans',
u'hs', u'ha', u'additionally', u'secondly', u'swag', u'stalkers', u'winded', u'jackoff',
u'criminals', u'athletes', u'wierd', u'crew', u'chatty', u'gents', u'downright', u'aids',
u'bodywork', u'fatty', u'sluts', u'favs', u'fuckin', u'dos', u'colorful', u'juliet', u'sorta', u'orgy',
u'omg', u'ether', u'gracias', u'ii', u'kim', u'unicorn', u'yep', u'unnecessary', u'www', u'keeper',
u'handicap', u'gaybi', u'stripping', u'overdue', u'cougars', u'dik', u'dig', u'bikers', u'perv',
u'aries', u'chemical', u'designer', u'bein', u'gamers', u'felonies', u'frequent', u'americans',
u'escort', u'gen', u'gem', u'affordable', u'jk', u'dater', u'sketchy', u'awesome', u'rancher',
u'ratchet', u'strait', u'obnoxious', u'inbetween', u'fatties', u'wifey', u'pups', u'pill',
u'meaning', u'downstairs', u'imma', u'squirter', u'strung', u'narrow', u'sadist', u'nester',

u'coat', u'eats', u'arguments', u'mmw', u'bulls', u'mmf', u'smoothe', u'needle', u'dddf',
u'contagious', u'lion', u'expert', u'underage', u'comptable', u'fur', u'pegging', u'loking',
u'darn', u'dishonesty', u'hates', u'glove', u'handjob', u'scale', u'nutshell', u'biggie',
u'raceethnicity', u'playin', u'stripper', u'loser', u'shemale', u'chasers', u'packing',
u'cocksuckers', u'fooling', u'high', u'dat', u'bromance', u'tranny', u'frank', u'babys',
u'curvey', u'disappointments', u'alaskan', u'dominatrix', u'superior', u'busted', u'proposal',
u'ridden', u'helper', u'oops', u'hippies', u'fyi', u'flaky', u'functional', u'chronic', u'ivory',
u'bum', u'cpl', u'partially', u'dangerous', u'daddydom', u'hating', u'mww', u'obsessive',
u'smokin', u'offs', u'visible', u'muse', u'mf', u'kewl', u'swingers', u'quickies', u'vocal',
u'educator', u'snob', u'rider', u'ocd', u'amazon', u'bottomvers', u'aquarius', u'faced',
u'ddfneg', u'beef', u'definetly', u'exterior', u'charlie', u'appearing', u'hunk', u'stereotype',
u'tht', u'psychos', u'gig', u'nerds', u'cos', u'ns', u'nt', u'gravitate', u'substitute', u'goer',
u'witch', u'swallower', u'geeks', u'yup', u'plastic', u'looin', u'screwing', u'population',
u'pro', u'mater', u'roleplaying', u'visitors', u'fck', u'mormon', u'bareback', u'freaking',
u'definite', u'pos', u'dis', u'tiger', u'U0001f61c', u'flakey', u'hipster', u'op', u'relocation',
u'sleazy', u'incorrect', u'curve', u'regulars', u'fag', u'rats', u'musts', u'attitudes',
u'supposedly', u'mmm', u'minot', u'waxed', u'penpal', u'goat', u'scientist', u'xb7', u'gud',
u'peeps', u'cookie', u'fuller', u'prejudice', u'wld', u'urinal', u'subslave', u'adonis', u'risky',
u'cleaner', u'font', u'cuck', u'mfm', u'bfgf', u'pr', u'flight', u'buck', u'cumslut', u'wheres',
u'starved', u'bimarrried', u'unstable', u'clown', u'nigga', u'kissers', u'asain', u'realtionship',
u'impatient', u'mainstream', u'attire', u'welcum', u'posing', u'thirsty', u'syndrome', u'dwn',
u'hum', u'beauties', u'yada', u'bois', u'abdl', u'ticklish', u'highschool', u'craziness', u'combo',
u'workin', u'closet', u'crude', u'jon', u'realtionship', u'dressers', u'stunning', u'vape',
u'definately', u'beast', u'whale', u'boot', u'owl', u'significantly', u'kidding', u'breeding',
u'bitches', u'granny', u'percent', u'boob', u'truckers', u'medications', u'sleeves']

2 : POSITIVE EMOTION WORDS

[u'hers', u'brings', u'circumstances', u'pursue', u'hidden', u'encourage', u'willingness', u'result', u'life', u'trusted', u'talents', u'fullest', u'world', u'memories', u'anothers', u'complete', u'learn', u'granted', u'compliment', u'levels', u'experiencing', u'challenge', u'allow', u'chosen', u'spoiling', u'earn', u'bring', u'expand', u'count', u'recognize', u'pursuit', u'motivate', u'devote', u'deepst', u'families', u'theirs', u'promises', u'reality', u'zone', u'discover', u'shared', u'selves', u'challenges', u'accomplish', u'space', u'embrace', u'valuable', u'rules', u'passions', u'journey', u'help', u'finer', u'safety', u'experiences', u'pace', u'lacking', u'towards', u'fears', u'requires', u'nurture', u'intentions', u'achieve', u'joys', u'interfere', u'precious', u'minds', u'secrets', u'mistakes', u'freedom', u'void', u'affect', u'enhance', u'spouses', u'faults', u'curiosity', u'overcome', u'downs', u'content', u'opportunity', u'thoughts', u'kinks', u'feelings', u'dreams', u'elses', u'sharing', u'effort', u'views', u'aspirations', u'order', u'them', u'fill', u'independence', u'perspective', u'fantasies', u'remain', u'advice', u'existence', u'significant', u'obligations', u'express', u'courage', u'desires', u'create', u'sexuality', u'purpose', u'ourselves', u'decision', u'others', u'imperfections', u'strive', u'interaction', u'concerns', u'knowledge', u'intensity', u'fulfilled', u'arrangements', u'alive', u'loneliness', u'path', u'ultimate', u'priorities', u'situations', u'improve', u'confide', u'greatest', u'hearts', u'apart', u'gift', u'frustrations', u'escape', u'everything', u'ensure', u'efforts', u'presence', u'puzzle', u'heal', u'focus', u'comfort', u'ideas', u'strengths', u'ways', u'lifes', u'direction', u'persons', u'changing', u'actions', u'separate', u'invest', u'decisions', u'opportunities', u'allows', u'options', u'members', u'changes', u'pray', u'parts', u'interact', u'motivation', u'creating', u'reasons', u'peoples',

u'peace', u'problems', u'helping', u'happier', u'vice', u'wildest', u'worlds', u'our', u'priority',
u'their', u'gain', u'joy', u'present', u'choices', u'happiness', u'other', u'urges']

3 : LOCATIONS

[u'broward', u'sturgis', u'china', u'rt', u'atl', u'champaign', u'new', u'traverse',
u'radius', u'canyon', u'union', u'stadium', u'ann', u'birmingham', u'creek', u'fox', u'marshall',
u'bloomington', u'cincinnati', u'burlington', u'myers', u'stockton', u'st', u'sb', u'haute',
u'square', u'sac', u'jefferson', u'federal', u'atlantic', u'southside', u'twin', u'locally', u'around',
u'boise', u'exit', u'october', u'ac', u'ar', u'az', u'monterey', u'midwest', u'baton', u'sarasota',
u'states', u'station', u'tri', u'england', u'highway', u'arlington', u'saint', u'walmart',
u'prescott', u'yakima', u'omaha', u'branson', u'towns', u'spring', u'victoria', u'stuart',
u'clinton', u'apartments', u'tally', u'uk', u'lane', u'land', u'watertown', u'northside',
u'hampton', u'street', u'tulsa', u'harbor', u'lee', u'parkway', u'bakersfield', u'eastside',
u'tahoe', u'blvd', u'blocks', u'upstate', u'harrisburg', u'haven', u'mills', u'pete', u'manchester',
u'rouge', u'jax', u'wilmington', u'butte', u'co', u'cc', u'acres', u'ct', u'midland', u'resort',
u'maui', u'paul', u'counties', u'massachusetts', u'inn', u'rural', u'india', u'modesto', u'japan',
u'rocky', u'reasonable', u'danville', u'honolulu', u'december', u'rome', u'germany',
u'renting', u'greenville', u'lexington', u'wa', u'wi', u'wv', u'cod', u'lauderdale', u'wheel',
u'vermont', u'jose', u'hills', u'peninsula', u'cruces', u'southwest', u'upper', u'franklin',
u'entrance', u'cal', u'islands', u'flint', u'fargo', u'marina', u'maine', u'lansing', u'interstate',
u'yuma', u'resides', u'daytona', u'cottage', u'suites', u'wyoming', u'streets', u'nebraska',
u'humid', u'cruz', u'march', u'duluth', u'transplant', u'newer', u'russian', u'manhattan',
u'fayetteville', u'gr', u'suite', u'barbara', u'collins', u'cloud', u'HUDSON', u'grove', u'triangle',

u'armor', u'border', u'indy', u'dodge', u'resident', u'northwest', u'riverside', u'cleveland',
u'delaware', u'cocoa', u'mississippi', u'hr', u'erie', u'lakeland', u'aug', u'panama', u'asia',
u'spokane', u'helena', u'redding', u'navy', u'hollywood', u'mins', u'retire', u'nevada',
u'northeast', u'lancaster', u'sail', u'slo', u'slc', u'il', u'belle', u'sacramento', u'clarksville',
u'amarillo', u'vista', u'jones', u'mile', u'mill', u'ma', u'hotels', u'france', u'fla', u'arbor',
u'lubbock', u'pine', u'johnson', u'canada', u'living', u'keith', u'midtown', u'ventura', u'burbs',
u'tuscaloosa', u'canton', u'clearwater', u'ocala', u'across', u'jc', u'middle', u'tour',
u'binghamton', u'topeka', u'uptown', u'acre', u'rapids', u'commute', u'syracuse', u'pueblo',
u'neighborhood', u'abq', u'airport', u'africa', u'asheville', u'destin', u'del', u'greeley', u'socal',
u'walla', u'okc', u'oxford', u'heights', u'morgantown', u'corpus', u'abilene', u'dayton', u'kc',
u'ks', u'boulder', u'toledo', u'century', u'monroe', u'keys', u'nola', u'district', u'ontario',
u'gainesville', u'desert', u'maryland', u'visalia', u'presently', u'knight', u'anywhere',
u'minnesota', u'bluff', u'billings', u'junction', u'shreveport', u'philadelphia', u'montgomery',
u'frederick', u'charles', u'national', u'stationed', u'homes', u'augusta', u'center', u'chico',
u'louisville', u'conway', u'ridge', u'waco', u'laredo', u'bus', u'brand', u'campus', u'paris',
u'pride', u'biloxi', u'hospital', u'mo', u'mn', u'mt', u'gate', u'mesa', u'london', u'oakland',
u'eugene', u'missoula', u'detroit', u'newport', u'oahu', u'pensacola', u'westside', u'bridge',
u'rogers', u'southeast', u'decatur', u'sept', u'anderson', u'nm', u'ne', u'marion', u'united',
u'athens', u'cheyenne', u'meridian', u'academy', u'kent', u'albuquerque', u'raleigh',
u'charleston', u'george', u'worcester', u'hoover', u'gulf', u'knoxville', u'region', u'annapolis',
u'nh', u'pennsylvania', u'ranch', u'eureka', u'pacific', u'caribbean', u'nudist', u'oc',
u'huntsville', u'hwy', u'finals', u'dairy', u'bozeman', u'lawrence', u'roanoke', u'florence',
u'oak', u'tallahassee', u'september', u'mission', u'milwaukee', u'avenue', u'fairbanks',
u'stephen', u'residing', u'medford', u'hill', u'paradise', u'humboldt', u'arcata', u'rochester',
u'wichita', u'village', u'martin', u'plaza', u'summertime', u'trucker', u'rainbow', u'lafayette',

u'buffalo', u'bellingham', u'anchorage', u'hilton', u'peoria', u'rockford', u'april', u'naples',
u'ave', u'rapid', u'smith', u'branch', u'suburbs']

4 : SEXUAL TERMS, NEGATIVE

[u'electricity', u'screaming', u'snuggles', u'increasing', u'locked', u'wink',
u'bringing', u'whim', u'spinning', u'arousal', u'eagerly', u'brains', u'bliss', u'plate',
u'goodnight', u'spin', u'remembering', u'ease', u'shadow', u'remind', u'constantly', u'inside',
u'crossdress', u'flood', u'stays', u'tear', u'sigh', u'melt', u'falling', u'degrade', u'savor',
u'yearn', u'troubles', u'buried', u'aside', u'tears', u'aroused', u'lovingly', u'raging',
u'squeezed', u'edges', u'burning', u'her', u'hed', u'exhausted', u'watches', u'watched',
u'pedestal', u'cums', u'someones', u'mask', u'displays', u'clock', u'fucks', u'grope', u'smiling',
u'licks', u'stoned', u'everytime', u'upstairs', u' imagine', u'sensations', u'body's',
u'remembers', u'magical', u'degraded', u'begins', u'wholl', u'worthless', u'awake', u'crying',
u'twinkle', u'opens', u'walked', u'scratch', u'continues', u'continued', u'wipe', u'gettin',
u'touches', u'pour', u'pieces', u'holds', u'cage', u'rising', u'beneath', u'wander', u'fantasys',
u'blows', u'blew', u'vise', u'wiggle', u'staring', u'sits', u'digging', u'dust', u'bulge', u'thier',
u'literally', u'drift', u'placed', u'endings', u'glance', u'backwards', u'quiver', u'stimulate',
u'strangers', u'delight', u'mercy', u'toast', u'draws', u'heals', u'excite', u'sting', u'kept',
u'drip', u'tha', u'victim', u'belongs', u'denied', u'beating', u'faint', u'fight', u'hump',
u'rhythm', u'blast', u'helpless', u'penetrated', u'buttons', u'masturbating', u'undivided',
u'undressed', u'increase', u'tingle', u'sheer', u'slam', u'ground', u'eachothers', u'burst',
u'moans', u'madness', u'punished', u'handled', u'exposed', u'heated', u'stare', u'groan',
u'sadness', u'poop', u'peak', u'pointing', u'heaven', u'darkness', u'depend', u'pushed', u'gasp',

u'groping', u'moaning', u'erecton', u'base', u'heartbeat', u'blush', u'unexpected', u'empty',
u'madly', u'cumming', u'showing', u'chairs', u'softness', u'nerves', u'ache', u'shine', u'rolls',
u'caresses', u'nap', u'picked', u'poke', u'wallet', u'popped', u'crossed', u'drops', u'fart',
u'outer', u'shaking', u'suckling', u'beside', u'overwhelming', u'cravings', u'torment',
u'brighten', u'brighter', u'fuel', u'sht', u'hits', u'invisible', u'jessy', u'senses', u'kicks',
u'periods', u'suckle', u'surface', u'began', u'shed', u'liking', u'piece', u'display', u'beats',
u'using', u'tips', u'survive', u'expose', u'sensually', u'climax', u'unforgettable', u'frisky',
u'differently', u'stealing', u'sparkle', u'starts', u'bouncing', u'louder', u'hunger', u'stepping',
u'sneaking', u'decides', u'wrapping', u'puts', u'stresses', u'burn', u'yelling', u'vaginal',
u'bodies', u'toss', u'anticipation', u'filled', u'treasure', u'silence', u'surprising', u'fills',
u'bending', u'takin', u'enthusiasm', u'killing', u'resist', u'sudden', u'protein', u'pat',
u'blankets', u'grin', u'serves', u'facing', u'matching', u'seconds', u'giggle', u'roam', u'quietly',
u'spreading', u'distracted', u'strikes', u'grabs', u'stood', u'sides', u'movement', u'purse',
u'compare', u'stir', u'pent', u'backed', u'compliments', u'smallest', u'pissed', u'dances',
u'patiently', u'dropping', u'aches', u'seduce', u'excites', u'laughed', u'knock', u'mirror',
u'lesson', u'cure', u'lend', u'shock', u'privately', u'closing', u'boner', u'dolled', u'bind',
u'hugged', u'tickled', u'thrill', u'slipping', u'popping', u'afterward', u'reaches', u'reached',
u'grabbed', u'beat', u'calling', u'awaits', u'deliver', u'manhood', u'gaze', u'tense', u'sets',
u'inviting', u'reminds', u'genitals', u'pin', u'clits', u'guiding', u'leaves', u'breathe',
u'entertained', u'loosen', u'mouths', u'enter', u'fade', u'erotically', u'filling', u'evil', u'flame',
u'parting', u'weak', u'devour', u'cracked', u'thumbs', u'petting', u'envision', u'heat',
u'ruined', u'pleases', u'masturbate', u'guilty', u'motions', u'bound', u'underneath', u'motion',
u'butterflies', u'tingling', u'gives', u'seduced', u'reveal', u'arrival', u'delicate', u'walls',
u'reward', u'kneeling', u'mood', u'stops', u'fantasize', u'grasp', u'glimpse', u'deeply',
u'suddenly', u'breaking', u'monotony', u'thrown', u'throws', u'passes', u'comfy', u'relieved',

u'explodes', u'tongues', u'reach', u'react', u'his', u'woke', u'rocking', u'stolen', u'ecstasy',
u'stimulated', u'cheer', u'gradually', u'refuses', u'womens', u'plunge', u'memory',
u'repeatedly', u'backs', u'allowing', u'breathing', u'wished', u'spoon', u'notes', u'sensation',
u'fold', u'pushing', u'rabbit', u'tap', u'crawling', u'shell', u'darkest', u'penetrate', u'view',
u'closes', u'opening', u'commands', u'safely', u'warming', u'banging', u'virginity', u'cake']

5 : NOT CONSISTENT

[u'second', u'until', u'busy', u'planned', u'steps', u'memorable', u'avail', u'time',
u'gloomy', u'minute', u'alone', u'celebrate', u'arriving', u'rained', u'rest', u'daytimes',
u'appointment', u'hours', u'prior', u'plan', u'approaching', u'set', u'monthly', u'eve', u'next',
u'process', u'chilly', u'ends', u'wed', u'permits', u'bday', u'after', u'greet', u'break',
u'appointments', u'convenient', u'phase', u'hrs', u'away', u'flexible', u'thurs', u'notice',
u'month', u'overnight', u'saturdays', u'fathers', u'visit', u'visits', u'waking', u'started',
u'unwind', u'times', u'frequently', u'warmer', u'eves', u'permit', u'weeks', u'turning',
u'threw', u'fell', u'daily', u'finally', u'shift', u'thru', u'arrive', u'till', u'membership',
u'sleeping', u'everyday', u'extended', u'blessed', u'starting', u'oclock', u'hour', u'plans',
u'coming', u'through', u'beginning', u'throughout', u'lunchtime', u'yesterday', u'moment',
u'spent', u'per', u'entire', u'min', u'upcoming', u'tues', u'unable', u'regularly', u'routine',
u'taking', u'minutes', u'every', u'socialize', u'end', u'over', u'days', u'wedding', u'twice',
u'fridays', u'til', u'stressful', u'chapter', u'fourth', u'passing', u'schedules', u'on', u'longer',
u'once', u'leaving', u'planning', u'half']

6 : DRUGS

[u'tolerate', u'herb', u'addicted', u'addict', u'drunks', u'absolutely', u'abusive', u'favors', u'smokers', u'exs', u'moderation', u'cig', u'illegal', u'druggies', u'scene', u'chew', u'smoke', u'recreational', u'baggage', u'smoked', u'poppers', u'drinkers', u'condoms', u'rarely', u'pnp', u'bugs', u'seldom', u'cigs', u'alcoholic', u'allergic', u'excess', u'addictions', u'alcoholics', u'booze', u'druggie', u'user', u'minimal', u'friendly', u'drinker', u'smoking', u'pets', u'tolerance', u'smokes', u'smoker', u'jealousy', u'diseases', u'marijuana', u'none', u'socially', u'drama', u'criminal', u'partier', u'cigars', u'drugs', u'social', u'addicts', u'abusers', u'tobacco', u'require', u'partake', u'pot', u'hardly', u'violent', u'habits', u'breakers', u'okay', u'cigarettes', u'alcohol', u'cigarette', u'users', u'stds', u'ghetto', u'excessive', u'record']

7 : NEGATIVE EMOTION WORDS

[u'hurt', u'how', u'keeps', u'easier', u'bitter', u'feeling', u'corny', u'huh', u'doin', u'fail', u'argue', u'vain', u'much', u'people', u'soo', u'nicest', u'exist', u'angry', u'wondering', u'pretend', u'smarter', u'seriously', u'wrong', u'ridiculous', u'depressed', u'nicer', u'sad', u'say', u'thinks', u'act', u'apologize', u'forgot', u'careful', u'case', u'refuse', u'its', u'always', u'really', u'guess', u'knowing', u'whole', u'surprised', u'bc', u'dislike', u'odd', u'makes', u'doubt', u'realize', u'truth', u'surely', u'scare', u'painful', u'fact', u'meant', u'guessing', u'wtf', u'less', u'hiding', u'asks', u'theyd', u'doing', u'worth', u'haha', u'worst', u'realizing', u'they', u'rejected', u'wasted', u'miserable', u'tough', u'petty', u'mistake', u'theyre', u'bore', u'way', u'hotter', u'dying', u'soooo', u'appear', u'unsure', u'usually', u'embarrassing',

u'hardest', u'still', u'entirely', u'impossible', u'truthfully', u'that', u'blame', u'hurts', u'these',
u'thinking', u'gonna', u'nowadays', u'lol', u'lot', u'trying', u'sooo', u'hell', u'actually',
u'reason', u'terrible', u'terribly', u'wow', u'damn', u'else', u'anymore', u'belong', u'popular',
u'idk', u'noone', u'win', u'cause', u'completely', u'perfectly', u'exactly', u'bet', u'hopeful',
u'werent', u'she', u'tend', u'horrible', u'admit', u'quit', u'harder', u'strange', u'lose', u'shes',
u'couldnt', u'pointless', u'offensive', u'besides', u'counts', u'struggle', u'seem', u'tells', u'bat',
u'bad', u'said', u'personally', u'suppose', u'shame', u'make', u'yet', u'somehow', u'dead',
u'craiglist', u'cliche', u'understood', u'happening', u'dealing', u'adds', u'suspect', u'frankly',
u'everybody', u'claim', u'crazy', u'mad', u'think', u'anyone', u'were', u'enough', u'wonder',
u'harsh', u'totally', u'unusual', u'appears', u'again', u'alright', u'forget', u'sucks', u'hurting',
u'easily', u'hard', u'pathetic', u'supposed', u'obviously', u'most', u'nervous', u'intimidated',
u'alot', u'point', u'cuz', u'excited', u'often', u'bored', u'putting', u'uncomfortable', u'obvious',
u'lame', u'warned', u'dreaming', u'embarrassed', u'nobody', u'unlikely', u'apparently',
u'sexier', u'normally', u'exists', u'joke', u'define', u'plain', u'frustrating', u'but', u'claiming',
u'hate', u'hide', u'ventured', u'keep', u'attract', u'badly', u'insecure', u'heck', u'rant',
u'though', u'approach', u'screw', u'dishonest', u'why', u'stuck', u'upset', u'annoying',
u'confused', u'weird', u'awkward', u'lonely', u'unfortunately', u'frustrated', u'fails',
u'myself', u'kill', u'seems', u'complain', u'there', u'odds', u'worse', u'far', u'awful', u'unlike',
u'what', u'trouble', u'needless', u'forgive', u'story', u'nothing', u'lying', u'dumb', u'hesitant',
u'try', u'anybody', u'skeptical', u'probably', u'settling', u'impress', u'honestly', u'category',
u'hence', u'boring', u'deny', u'ones', u'difficult', u'because', u'scared', u'scares', u'rare',
u'swear', u'conclusion', u'kinda']

8 : POSITIVE PERSONAL CHARACTERISTICS

[u'natured', u'successful', u'funny', u'person', u'responsible', u'creative', u'witty', u'respectable', u'ambitious', u'assertive', u'understanding', u'positive', u'spoken', u'open', u'touchy', u'mature', u'caring', u'grounded', u'bubbly', u'fiercely', u'sense', u'attitude', u'brutally', u'relaxed', u'energetic', u'intelligent', u'classy', u'openminded', u'chivalrous', u'trusting', u'gentle', u'listener', u'optimistic', u'generous', u'mannered', u'secure', u'kindhearted', u'cuddly', u'truthful', u'conversationalist', u'judgmental', u'personality', u'gentleman', u'easy', u'independent', u'humble', u'humor', u'quirky', u'polite', u'loyal', u'sociable', u'dedicated', u'playful', u'honest', u'talented', u'charismatic', u'laidback', u'thoughtful', u'very', u'minded', u'spontaneous', u'motivated', u'feely', u'courteous', u'sensitive', u'smart', u'flirty', u'loving', u'willed', u'compassionate', u'charming', u'loveable', u'overall', u'sassy', u'shy', u'hearted', u'sarcastic', u'sweet', u'articulate', u'kisser', u'judgemental', u'calm', u'silly', u'communicative', u'hardworking', u'outspoken', u'intellectual', u'down', u'genuinely', u'dependable', u'educated', u'sincere', u'passionate', u'careing', u'layed', u'sophisticated', u'romantic', u'straightforward', u'witted', u'outgoing', u'rounded', u'goofy', u'spirited', u'easygoing', u'extremely', u'genuine', u'humorous', u'patient', u'respectful', u'reserved', u'communicator', u'reliable', u'laugh', u'upbeat', u'talkative', u'laid', u'hopeless', u'lovable', u'oriented', u'earth', u'considerate', u'fault', u'confident', u'trustworthy', u'blunt', u'quiet', u'peaceful', u'attentive', u'spirit', u'sence', u'driven', u'faithful', u'empathetic', u'cultured', u'personable', u'devoted', u'mellow', u'protective', u'adventurous', u'kind', u'affectionate', u'orientated', u'artistic', u'stable', u'outdoorsy', u'nurturing', u'individual', u'sufficient']

9 : SEXUAL TERMS

[u'blowjobs', u'lots', u'massaging', u'masturbation', u'being', u'stoking', u'bjs', u'action', u'hugs', u'shower', u'pda', u'kissing', u'blowing', u'snuggling', u'jerking', u'nip', u'topped', u'holding', u'fingered', u'receiving', u'topping', u'rubs', u'anal', u'swallowing', u'eating', u'showering', u'grinding', u'hj', u'kisses', u'makeout', u'fingering', u'nipple', u'swapping', u'making', u'touching', u'foreplay', u'messages', u'jo', u'baths', u'fondling', u'laughing', u'bottoming', u'rubbing', u'showers', u'hugging', u'pleasing', u'giving', u'cuddling', u'teasing', u'recieving', u'oral', u'licking', u'jacking', u'cuddles', u'caressing', u'rimmed', u'sucking', u'fucking', u'edging', u'throating', u'steamy', u'flirting', u'rimming', u'mutual']

10 : DATES & LEISURE ACTIVITIES

[u'hanging', u'glass', u'clubbing', u'wine', u'breakfast', u'homebody', u'bonfire', u'meal', u'bowl', u'movie', u'drunk', u'partying', u'netflix', u'meals', u'stuff', u'gamble', u'pizza', u'coffee', u'relaxing', u'diner', u'together', u'dance', u'club', u'cold', u'hookah', u'grill', u'occasions', u'drink', u'go', u'grab', u'party', u'laughs', u'gym', u'bar', u'potato', u'lunch', u'shop', u'cocktails', u'dates', u'bottle', u'cooked', u'exercise', u'occasionally', u'occasion', u'buy', u'binge', u'beer', u'catch', u'workout', u'clubs', u'ice', u'dinner', u'cook', u'drinks', u'parties', u'beers', u'food', u'conversations', u'nails', u'tea', u'occasionally', u'dine', u'bars', u'drinking', u'out', u'cocktail', u'conversation']

11 : PAST VERBS OF EXPERIENCE

[u'dreamed', u'came', u'worked', u'craigslist', u'had', u'brought', u'years', u'quite', u'turned', u'saw', u'knew', u'many', u'ive', u'personals', u'went', u'luck', u'ago', u'gone', u'past', u'postings', u'found', u'missed', u'failed', u'last', u'acted', u'posted', u'seemed', u'realized', u'craving', u'cl', u'theyve', u'gained', u'lately', u'lasted', u'told', u'stressed', u'hooked', u'created', u'enjoyed', u'months', u'was', u'cheated', u'clue', u'heard', u'now', u'caught', u'discovered', u'since', u'awhile', u'learned', u'tried', u'stopped', u'miss', u'wasnt', u'forgotten', u'liked', u'used', u'success', u'dull', u'lost', u'previously', u'noticed', u'ever', u'never', u'met', u'waited', u'before', u'seen', u'decided', u'iv', u'left', u'lied', u'felt', u'died', u'did', u'talked', u'earlier', u'slept', u'dated', u'broken', u'done', u'gotten', u'veve', u'burned', u'encountered', u'bottomed', u'havent', u'didnt', u'fantasizing', u'sadly', u'received', u'almost', u'helped', u'thought', u'screwed', u'changed', u'got', u'wanted', u'took', u'been', u'several', u'gave', u'happened', u'given', u'youve', u'urge', u'hasnt', u'known', u'wondered', u'passed', u'numerous', u'recently', u'became', u'called', u'dealt', u'ended', u'thus', u'kissed', u'fantasized', u'looked', u'made']

12 : HOBBIES

[u'music', u'gaming', u'thrones', u'basketball', u'nature', u'country', u'oldies', u'films', u'manga', u'avid', u'punk', u'technology', u'hobbies', u'sci', u'tv', u'science', u'scifi', u'blues', u'listening', u'folk', u'animals', u'video', u'comedy', u'books', u'scary', u'hobby', u'politics', u'drawing', u'indie', u'techno', u'dogs', u'musical', u'rpgs', u'board', u'anime', u'songs', u'shows', u'fi', u'learning', u'cats', u'comics', u'nerd', u'horror', u'history', u'culture', u'game', u'draw', u'plays', u'poetry', u'variety', u'stories', u'star', u'wars',

u'comedies', u'reader', u'sorts', u'classic', u'rock', u'reading', u'among', u'doctor', u'drums',
u'film', u'genres', u'arts', u'pop', u'alternative', u'martial', u'genre', u'sing', u'photography',
u'novels', u'classical', u'rampb', u'documentaries', u'writing', u'fiction', u'fashion', u'rap',
u'metal', u'geek', u'gamer', u'cultural', u'xbox', u'nerdy', u'football', u'geeky', u'comic',
u'jazz', u'guitar', u'cultures', u'fan', u'computers', u'hip', u'cartoons', u'art', u'baseball',
u'bluegrass', u'literature', u'sports', u'videogames', u'piano', u'kinds']

13 : SEX AND BODY TERMS

[u'yellow', u'shaving', u'wooden', u'blouse', u'calves', u'sheet', u'exposing', u'crawl',
u'deepthroating', u'bomb', u'mens', u'lube', u'pits', u'smash', u'cunt', u'hat', u'bred', u'shoots',
u'roll', u'rolling', u'sneakers', u'shirts', u'squirting', u'tightly', u'everywhere', u'thumb',
u'massive', u'stuffed', u'cherry', u'strapon', u'busting', u'stretch', u'reflex', u'edged',
u'vacuum', u'gliding', u'butthole', u'cream', u'pearl', u'tummy', u'condom', u'sleeve', u'belt',
u'nylons', u'suit', u'holes', u'sweats', u'flowing', u'whipping', u'drill', u'pair', u'highs',
u'scent', u'shorts', u'steel', u'seductive', u'tape', u'lotion', u'ripe', u'ball', u'crack', u'loaded',
u'erect', u'lift', u'stinky', u'tickle', u'bra', u'plow', u'boobies', u'wore', u'worn', u'missionary',
u'sack', u'lifted', u'smacking', u'dildo', u'da', u'air', u'foreskin', u'tribbing', u'tounge', u'flash',
u'cap', u'clothing', u'em', u'worshipped', u'thong', u'ring', u'fours', u'rear', u'throw', u'jacked',
u'moves', u'milking', u'undies', u'lip', u'flesh', u'perfume', u'double', u'thongs', u'spray',
u'shiny', u'smelly', u'restrained', u'kitty', u'button', u'boots', u'hose', u'wig', u'sperm',
u'front', u'topless', u'boxers', u'candy', u'gagged', u'twist', u'clothes', u'bikini', u'pumping',
u'blindfold', u'shirt', u'vibrators', u'soaking', u'bone', u'cum', u'pussies', u'grease',
u'saggy', u'hood', u'bag', u'ears', u'precum', u'tank', u'butter', u'cocksucking', u'flops',

u'pantyhose', u'curling', u'diaper', u'covers', u'sticky', u'sling', u'jacket', u'yummy', u'string',
u'dim', u'dip', u'run', u'nylon', u'gloryhole', u'armpits', u'pink', u'strokes', u'shoes',
u'fingertips', u'monster', u'gagging', u'underwear', u'briefs', u'sheets', u'skull', u'stretched',
u'nose', u'bathe', u'bang', u'pissing', u'doggy', u'washing', u'pressing', u'vibrator',
u'worshiping', u'delicious', u'oil', u'fist', u'lotions', u'leash', u'fuk', u'sandals', u'pee',
u'fleshlight', u'brushing', u'worshipping', u'sink', u'limp', u'nips', u'prostate', u'ample',
u'vagina', u'optional', u'insert', u'glory', u'flip', u'pie', u'dressed', u'hats', u'feeding', u'rise',
u'shining', u'cloths', u'washed', u'outfits', u'top', u'faces', u'xxx', u'plug', u'bath', u'labia',
u'tit', u'navel', u'rod', u'swallowed', u'handles', u'teased', u'roses', u'eagle', u'latex', u'jizz',
u'naked', u'pole', u'mount', u'slippery', u'rose', u'towel', u'toilet', u'tasty', u'rubber', u'flick',
u'tee', u'suk', u'sticking', u'glide', u'stroked', u'cups', u'soapy', u'cleaned', u'windows',
u'doggie', u'slapped', u'outfit', u'cloth', u'suckin', u'wrestle', u'gang', u'duct', u'fed', u'straps',
u'covered', u'biting', u'pipe', u'chicken', u'cleavage', u'thrusting', u'cologne', u'draining',
u'rocket', u'sweat', u'faggot']

14 : ??? WHAT TO DO NEXT

[u'basics', u'dnt', u'wana', u'want', u'mabey', u'whit', u'shall', u'letting', u'and',
u'things', u'casually', u'intend', u'acquaintance', u'whoever', u'along', u'wherever', u'so',
u'course', u'take', u'thatll', u'at', u'to', u'ta', u'pick', u'more', u'itd', u'endlessly', u'beforehand',
u'also', u'sometimes', u'kno', u'knw', u'be', u'aswell', u'anything', u'up', u'anxious', u'let',
u'correspond', u'intersted', u'should', u'first', u'off', u'bro', u'somewhere', u'realy', u'do',
u'aim', u'alittle', u'mayb', u'wat', u'converse', u'tryna', u'somthing', u'brave', u'can', u'chit',
u'introduce', u'remotely', u'only', u'amd', u'amp', u'you'd', u'me', u'loose', u'ahead', u'hurry',

u'a', u'ya', u'lil', u'whatever', u'dialogue', u'intriguing', u'ready', u'gt', u'wit', u'meet', u'may',
u'timer', u'conversate', u'i', u'well', u'blo', u'even', u'specifics', u'give', u'better', u'frist', u'id',
u'meetup', u'just', u'somethings', u'shoot', u'snd', u'talk', u'seeing', u'sooner', u'fast', u'little',
u'gotta', u'show', u'get', u'kool', u'then', u'andor', u'abt', u'def', u'least', u'b', u'simply',
u'meeting', u'talking', u'convo', u'bit', u'back', u'invite', u'blaze', u'carplay', u'alil', u'll',
u'willing', u'have', u'unless', u'emailtext', u'nd', u'like', u'about', u'goin', u'would', u'gos',
u'too', u'us', u'n', u'able', u'than', u'someplace', u'afterwards', u'when', u'lets', u'or', u'buds',
u'jus', u'all', u'maby', u'rather', u'bout', u'might', u'hit', u'are', u'initially', u'wanna', u'sext',
u'few', u'something', u'anyhow', u'ahold', u'lemme', u'freinds', u'atleast', u'know', u'holler',
u'getting', u'whether']

15 : EMPLOYMENT, HOUSING & FINANCIAL SITUATION

[u'fix', u'afford', u'laundry', u'spot', u'drive', u'jobs', u'spare', u'legendary',
u'schedule', u'work', u'office', u'split', u'security', u'pays', u'cost', u'car', u'cat', u'cars',
u'roommates', u'steady', u'business', u'basement', u'manage', u'vehicle', u'privacy', u'truck',
u'license', u'home', u'room', u'roof', u'cleaning', u'transportation', u'crib', u'paid', u'pay',
u'money', u'drivers', u'house', u'insurance', u'gas', u'source', u'apt', u'epic', u'vehicles',
u'access', u'private', u'responsibilities', u'fixed', u'host', u'studio', u'property', u'rental',
u'plenty', u'expenses', u'place', u'rent', u'bought', u'paying', u'apartment', u'ticket', u'suv',
u'chores', u'condo', u'due', u'income', u'bills', u'job', u'employment', u'own']

16 : POSITIVE PERSONAL CHARACTERISTICS

[u'elegant', u'dork', u'organized', u'modest', u'unselfish', u'nympho', u'vary', u'dreamer', u'matured', u'wholesome', u'fairly', u'innocent', u'liberal', u'accomplished', u'love', u'logical', u'wicked', u'explorer', u'conscious', u'opinionated', u'workaholic', u'fashionable', u'bright', u'equipped', u'considers', u'joyful', u'lively', u'politically', u'as', u'unpredictable', u'authentic', u'descrete', u'determined', u'respectfull', u'formal', u'stylish', u'diverse', u'carrying', u'high', u'tallish', u'hygene', u'degreed', u'borderline', u'hilarious', u'smartass', u'travelled', u'behaved', u'adjusted', u'transparent', u'passable', u'stamina', u'desirable', u'spunky', u'highly', u'chilled', u'outdoorsman', u'selfless', u'passive', u'adorable', u'balanced', u'grateful', u'forgiving', u'thinker', u'productive', u'ethic', u'wealthy', u'opened', u'rational', u'insatiable', u'terrific', u'noble', u'knowledgeable', u'conservative', u'low', u'tendencies', u'vivacious', u'selective', u'exceptionally', u'musically', u'crafty', u'hyper', u'pretty', u'adores', u'protector', u'described', u'outstanding', u'maintenance', u'collected', u'bonus', u'banter', u'savvy', u'descreet', u'smells', u'intuitive', u'cheerful', u'focused', u'uptight', u'ridiculously', u'handy', u'keen', u'distinguished', u'fluent', u'faithfull', u'virgo', u'girlie', u'key', u'poet', u'both', u'eclectic', u'eccentric', u'headed', u'tempered', u'appreciative', u'atheist', u'perverted', u'active', u'side', u'loud', u'warped', u'twisted', u'somewhat', u'nonsmoking', u'ethical', u'inquisitive', u'versed', u'bold', u'refined', u'amazingly', u'temper', u'fashioned', u'generally', u'talker', u'androgynous', u'extrovert', u'personalities', u'speaking', u'comedian', u'bossy', u'hippy', u'possessive', u'maintained', u'sensible', u'dramatic', u'nonjudgmental', u'uncomplicated', u'presentable', u'practical', u'skilled', u'spicy', u'obsessed', u'carefree', u'controlling', u'stubborn', u'sharp', u'tolerant', u'goof', u'libido', u'encouraging', u'pleasant', u'disposition', u'insightful', u'charm', u'cynical', u'manly', u'dancer', u'timid', u'gestures', u'openly', u'an', u'\xe2\u2022', u'varied', u'imaginative', u'tidy', u'enthusiastic',

u'inspiring', u'exceptional', u'literate', u'seasoned', u'mildly', u'vibrant', u'smarts',
u'intellegent', u'introvert', u'wise', u'expressive', u'mischievous', u'feisty', u'settled',
u'excellent', u'demanding', u'philosophical', u'entertaining', u'adventuresome', u'complex',
u'scorpio', u'conversational', u'gifted', u'core', u'entrepreneur', u'seeker', u'loveing',
u'sporty', u'prep', u'secured', u'exhibitionist', u'funloving', u'realist', u'independant',
u'giver', u'adventerous', u'cuddler', u'feminist', u'jokes', u'flirtatious', u'dorky', u'halfway',
u'joking', u'provider', u'themselves', u'style', u'harmless', u'impeccable', u'hopelessly',
u'stoner', u'edgy', u'reliant', u'competitive', u'humour', u'tends', u'cautious', u'introverted',
u'freak', u'daring', u'goofball', u'streak', u'clever', u'inclined', u'tad', u'progressive',
u'honorable', u'pleaser', u'artsy', u'worldly', u'wild', u'heavily', u'incredibly', u'demeanor',
u'mined', u'leader', u'loner', u'intimidating']

17 : POSITIVE PHYSICAL CHARACTERISTICS

[u'blonde', u'piercings', u'cleancut', u'blu', u'salt', u'trimmed', u'haireyes', u'dark',
u'dimples', u'handsome', u'brunette', u'moderately', u'tats', u'blond', u'caramel', u'pds',
u'pale', u'bl', u'uc', u'glasses', u'cm', u'bald', u'irish', u'red', u'brn', u'pierced', u'length', u'br',
u'green', u'highlights', u'muscular', u'brown', u'freckles', u'curly', u'neatly', u'complexion',
u'pounds', u'toned', u'ft', u'auburn', u'grey', u'hair', u'wavy', u'pepper', u'graying',
u'bluegreen', u'sandy', u'beard', u'beefy', u'gray', u'broad', u'goatee', u'med', u'tatts',
u'ginger', u'reddish', u'bearded', u'tattooed', u'shoulder', u'skin', u'brownish', u'olive',
u'waist', u'tall', u'eyes', u'eyed', u'skinned', u'thinning', u'strawberry', u'tattoos', u'built',
u'build', u'weigh', u'scruffy', u'cut', u'mod', u'mustache', u'lb', u'hazel', u'tone', u'italian',

u'blue', u'haired', u'lbs', u'brbr', u'medium', u'trim', u'tanned', u'blondish', u'dreads', u'c',
u'gotee', u'tan', u'light', u'brwn', u'buzzed', u'inch', u'facial', u'lean']

18 : EVENING & LEISURE ACTIVITIES

[u'snuggled', u'thunder', u'secluded', u'wind', u'rv', u'dipping', u'passenger',
u'theaters', u'winter', u'spooning', u'alley', u'winters', u'lawn', u'canoe', u'lounging',
u'buying', u'beverage', u'sand', u'followed', u'door', u'waves', u'backyard', u'patio',
u'broadway', u'candlelight', u'getaway', u'church', u'hitting', u'fire', u'sight', u'nowhere',
u'chillin', u'scenery', u'popcorn', u'forest', u'stars', u'brew', u'jurassic', u'rooms', u'desk',
u'pier', u'garage', u'errands', u'spots', u'places', u'ins', u'blanket', u'tickets', u'salon', u'pjs',
u'rocks', u'storm', u'festival', u'luxury', u'runs', u'steak', u'steam', u'convertible', u'kitchen',
u'fest', u'thursdays', u'rain', u'soak', u'cooler', u'wal', u'mtn', u'market', u'candles',
u>window', u'hall', u'lunches', u'sunset', u'stroll', u'concert', u'breeze', u'moonlight',
u'starbucks', u'sundays', u'trailer', u'event', u'pub', u'sauna', u'grocery', u'randomly', u'cave',
u'lit', u'trees', u'workouts', u'throwing', u'bench', u'wings', u'pedicures', u'drag', u'dvd',
u'barbecue', u'nites', u'cruise', u'camper', u'parked', u'sipping', u'tent', u'kicking', u'brunch',
u'library', u'rodeo', u'cruising', u'waterfront', u'deck', u'picking', u'sunshine', u'flying',
u'fireworks', u'hangin', u'fly', u'dutch', u'cheese', u'track', u'corner', u'curled', u'storms',
u'tub', u'restroom', u'sights', u'strip', u'mart', u'chats', u'bookstore', u'youtube', u'remote',
u'talks', u'enjoying', u'sitting', u'destination', u'marathons', u'cozy', u'restaurant', u'flags',
u'candle', u'tripping', u'planet', u'cafe', u'dessert', u'climb', u'trail', u'snacks', u'summers',
u'cutting', u'toke', u'recipes', u'finest', u'cabin', u'tvmovies', u'dunes', u'pit',
u'thunderstorms', u'fireplace', u'nail', u'jumping', u'jacuzzi', u'driving', u'soda', u'hopping',

u'burger', u'classes', u'cookout', u'radio', u'bingo', u'bathroom', u'cycle', u'christmas', u'trip',
u'setting', u'clouds', u'spur', u'appetizers', u'catching', u'sunrise', u'hulu', u'swing',
u'convention', u'tanning', u'marathon', u'surf', u'wood', u'picnic', u'parking', u'cafes',
u'barefoot', u'moon', u'sofa', u'nightlife', u'grass', u'grabbing', u'shore', u'sale', u'redbox',
u'fancy', u'store', u'watchin', u'sky', u'holidays', u'aways', u'porch', u'spa', u'sip', u'lounge',
u'outings', u'jog', u'walk', u'table', u'campfire', u'boardwalk']

19 FETISH TERMS

[u'handcuffs', u'skirts', u'bottom', u'kinky', u'role', u'servicing', u'dirty', u'diapers',
u'dress', u'scat', u'torture', u'choking', u'sampm', u'bb', u'tied', u'rough', u'cd', u'denial',
u'penetration', u'scenes', u'forced', u'fetishes', u'ws', u'wax', u'leather', u'heels', u'bondage',
u'facials', u'verbal', u'mild', u'toys', u'piss', u'rope', u'plugs', u'taboo', u'ons', u'blindfolds',
u'fetish', u'bds', u'panties', u'wearing', u'slutty', u'makeup', u'watersports', u'pain', u'gags',
u'stockings', u'hardcore', u'slapping', u'clamps', u'stretching', u'restraints', u'blood',
u'extreme', u'tickling', u'nasty', u'lingerie', u'dildos', u'pig', u'strap', u'dressing', u'whips',
u'raw', u'spanked', u'porn', u'chastity', u'multiple', u'cbt', u'domination', u'kink', u'fisting',
u'rape', u'abuse', u'spankings', u'spanking', u'humiliation', u'panty', u'cross', u'roleplay']

20 BODY PARTS & SEX PRACTICES

[u'crotch', u'fingers', u'whip', u'wash', u'legs', u'mouth', u'scream', u'spit', u'lubed',
u'wetness', u'floor', u'push', u'chair', u'clitoris', u'swollen', u'pulled', u'knee', u'breath',

u'pull', u'caress', u'nibble', u'pressed', u'bounce', u'unzip', u'leg', u'nibbling', u'bent', u'lock',
u'bend', u'slip', u'thigh', u'shut', u'thrust', u'gently', u'nutt', u'sliding', u'throbbing',
u'stomach', u'hand', u'juices', u'chin', u'slap', u'finish', u'squirm', u'unlocked', u'forcing',
u'firmly', u'juice', u'hips', u'undress', u'lips', u'pounding', u'rub', u'caressed', u'cheeks',
u'pulling', u'skirt', u'hands', u'slide', u'ankles', u'yonis', u'wrapped', u'neck', u'beg', u'bare',
u'lightly', u'passionately', u'sniff', u'slowly', u'shoulders', u'tongue', u'inner', u'smack',
u'force', u'lights', u'arms', u'warm', u'seed', u'massaged', u'softly', u'bury', u'shake', u'lap',
u'seat', u'grip', u'tease', u'remove', u'thighs', u'blindfolded', u'fondle', u'wrap', u'socks',
u'spread', u'whisper', u'aching', u'forehead', u'explode', u'onto', u'toe', u'ram', u'finger',
u'ear', u'choke', u'head', u'tip', u'tie', u'shove', u'gag', u'begging', u'shaft', u'rubbed', u'toes',
u'feet', u'pump', u'moist', u'dump', u'arm', u'sweep', u'rip', u'lower', u'cheek', u'edge',
u'squeezing', u'moan', u'sore', u'dripping', u'slides', u'knees', u'kneel', u'pants', u'wall',
u'press', u'brush', u'squeeze']

21 STDS & PREGNANCY

[u'dnd', u'vasectomy', u'heightweight', u'financially', u'hygienic', u'ddf', u'disease',
u'sane', u'tested', u'freshly', u'ub', u'fresh', u'discret', u'drug', u'ultra', u'neat', u'stdhiv',
u'negative', u'proportionate', u'reasonably', u'unattached', u'df', u'dd', u'ddfree', u'desease',
u'showered', u'non', u'undetectable', u'std', u'squeaky', u'mobile', u'clean', u'drugdisease',
u'popper', u'hivstd', u'neg', u'shaven', u'shaved', u'ddfhiv', u'hygiene', u'dampd',
u'diseasedrug', u'relatively', u'ddd', u'nonsmoker', u'disease', u'bug', u'must', u'free',
u'sober', u'healthy', u'decently', u'hwp', u'expect', u'discrete', u'hiv', u'groomed']

22 PHYSICAL CHARACTERISTICS

[u'verstop', u'pubic', u'silver', u'averageathletic', u'nice', u'scruff', u'runner', u'bodybuilder', u'htwt', u'dicked', u'nicely', u'player', u'curvaceous', u'azz', u'smelling', u'stature', u'sz', u'natural', u'ss', u'sp', u'sm', u'bush', u'saltpepper', u'prefered', u'backside', u'padding', u'teddy', u'tatoos', u'blah', u'blondes', u'jeans', u'dds', u'package', u'shapely', u'brazilian', u'darker', u'af', u'dyed', u'smile', u'tatted', u'gorgeous', u'sucker', u'silky', u'hangers', u'hott', u'heads', u'bimwm', u'body', u'firm', u'chest', u'd', u'slimathletic', u'shooters', u'breasted', u'defined', u'thugs', u'partial', u'mediterranean', u'fifteen', u'standing', u'heavysset', u'extraordinarily', u'curvythick', u'fivenine', u'ftin', u'butts', u'tattoo', u'cleanddf', u'swimmers', u'naturally', u'versbottom', u'ink', u'brnbrn', u'greek', u'approximately', u'emo', u'wt', u'redneck', u'hairless', u'gut', u'uniform', u'ibs', u'shooter', u'approx', u'haircut', u'redbone', u'inked', u'shade', u'german', u'proportion', u'lite', u'whitehispanic', u't', u'brunettes', u'thickness', u'colored', u'circumcised', u'pack', u'heritage', u'lknng', u'slimfit', u'fuzzy', u'physique', u'redheads', u'runners', u'hw', u'ht', u'athletically', u'mex', u'piercing', u'slimskinny', u'weakness', u'features', u'fluffy', u'wears', u'tiny', u'mushroom', u'belly', u'athleticmuscular', u'builds', u'teeth', u'round', u'shave', u'cummers', u'french', u'cummer', u'bod', u'pluses', u'blooded', u'jock', u'killer', u'goth', u'exotic', u'cuban', u'brownskin', u'plump', u'endowed', u'complected', u'abs', u'surfer', u'dampdf', u'muscles', u'muscled', u'roughly', u'cheded', u'giant', u'brownblue', u'xtra', u'cup', u'semi', u'hippie', u'framed', u'moderate', u'boned', u'boyish', u'bear', u'swimmer', u'multi', u'builder', u'muscle', u'soft', u'islander', u'slight', u'preppy', u'dresses', u'dresser', u'lightskin', u'meaty', u'musc', u'carmel', u'burly', u'tom', u'tool', u'rat', u'jean', u'wear', u'pubes', u'furry', u'hung', u'brbl', u'balding', u'chocolate', u'cowboy', u'huge', u'buzz',

u'virgin', u'wide', u'lighter', u'twinkish', u'bodied', u'polish', u'negddf', u'blondblue', u'flat',
u'weighs', u'buff', u'beards', u'shaped', u'prop', u'grn', u'bearish', u'shades', u'penis',
u'puertorican', u'accent', u'cubs', u'bellies', u'linebacker', u'weighing', u'browngreen',
u'thug', u'collar', u'rocker', u'endowment', u'skater', u'ripped', u'perky', u'below']

23 COMMUNICATION DEVICES & CONTACT INFO

[u'yahoo', u'emailing', u'address', u'unsolicited', u'verification', u'pics', u'via',
u'trade', u'link', u'chat', u'exchanges', u'email', u'vv', u'verify', u'forth', u'webcam',
u'exchanged', u'voice', u'mails', u'sending', u'cam', u'speed', u'snapchat', u'instagram', u'fb',
u'pictures', u'skype', u'messaging', u'e', u'facebook', u'cell', u'addresses', u'links', u'emails',
u'endless', u'username', u'swap', u'trading', u'messages', u'direct', u'mailing', u'kik', u'mail',
u'texting', u'numbers', u'snap', u'exchange', u'account', u'messenger', u'text', u'chatting',
u'phone', u'contact', u'facetime', u'faster', u'blocked', u'exchanging', u'calls', u'texts',
u'number', u'app', u'tag']

24 LEISURE ACTIVITIES

[u'hop', u'atv', u'roller', u'atvs', u'ballet', u'bicycles', u'nascar', u'roadtrips',
u'surfing', u'tasting', u'scuba', u'farmers', u'going', u'hockey', u'boat', u'diving', u'yoga',
u'puzzles', u'pulls', u'volleyball', u'resorts', u'travelling', u'jet', u'darts', u'cruises',
u'wheeler', u'projects', u'sightseeing', u'fixing', u'garden', u'paintball', u'motorcycle',
u'scenic', u'socializing', u'painting', u'bikes', u'chess', u'roads', u'gardens', u'boats', u'target',

u'historical', u'muddin', u'antique', u'wineries', u'float', u'rodeos', u'harleys', u'lifting',
u'backpacking', u'roading', u'television', u'cycling', u'floating', u'sales', u'gambling',
u'crafts', u'bass', u'dirt', u'horse', u'performances', u'exercising', u'lakes', u'singing',
u'tennis', u'ride', u'ect', u'indoor', u'trucks', u'motorcycling', u'frisbee', u'museum',
u'rafting', u'foods', u'boarding', u'mini', u'dog', u'disney', u'putt', u'sushi', u'whiskey',
u'kayak', u'sailing', u'tractor', u'theatre', u'climbing', u'venues', u'outs', u'crafting', u'road',
u'rollerblading', u'galleries', u'golfing', u'running', u'quads', u'paint', u'stargazing', u'ufc',
u'lounges', u'sunrises', u'musicals', u'rivers', u'paddle', u'boxing', u'tubing', u'gazing',
u'yard', u'stores', u'pools', u'skating', u'shooting', u'bicycle', u'campfires', u'coasters',
u'jogging', u'poker', u'harley', u'theme', u'antiques', u'snow', u'mud', u'bbqing', u'craft',
u'bands', u'hunt', u'zoos', u'scrabble', u'bake', u'water', u'opera', u'exploring', u'cards',
u'snowmobiling', u'playing', u'outside', u'shops', u'drives', u'jeep', u'cinema', u'bicycling',
u'disc', u'racing', u'guns', u'thrift', u'miniature', u'horses', u'malls', u'softball', u'soccer',
u'snowboarding', u'mountain', u'volunteering', u'wrestling', u'motor', u'nfl', u'snorkeling',
u'dive', u'trivia', u'auctions', u'ski']

25 NOT CONSISTENT

[u'funk', u'specialist', u'lgbt', u'disability', u'unit', u'household', u'counseling',
u'damage', u'schools', u'xs', u'rn', u'millions', u'project', u'episode', u'released', u'patch',
u'disaster', u'hammer', u'havnt', u'accident', u'hated', u'pregnancy', u'academic', u'sleeps',
u'caregiver', u'prison', u'attorney', u'losing', u'dollars', u'becuase', u'childs', u'embark',
u'collage', u'yards', u'studied', u'studies', u'stole', u'applications', u'decades', u'nation',
u'beloved', u'grandma', u'saving', u'vote', u'fooled', u'prospect', u'illness', u'prime',

u'surgery', u'concentrate', u'allowance', u'formed', u'costa', u'costs', u'traffic', u'cannabis',
u'neighbor', u'homeless', u'puppy', u'mass', u'broaden', u'saved', u'joined', u'nurse',
u'november', u'healthier', u'coolest', u'marriages', u'sitter', u'plane', u'roots', u'mistreated',
u'courses', u'owners', u'castle', u'guest', u'focusing', u'\xe2', u'despise', u'preparing',
u'spirits', u'roommate', u'thou', u'bachelor', u'hollow', u'devil', u'champion', u'budget',
u'flooded', u'upbringing', u'concrete', u'unto', u'readers', u'shitty', u'suffering', u'temple',
u'occupy', u'nieces', u'faded', u'consumed', u'graduating', u'technical', u'cousins',
u'practicing', u'rode', u'hormones', u'counter', u'meantime', u'bump', u'mainland', u'scars',
u'blown', u'system', u'depressing', u'allot', u'decade', u'jan', u'temporarily', u'packed',
u'drum', u'cousin', u'government', u'haul', u'replacement', u'biology', u'housing', u'pension',
u'oldest', u'toll', u'officially', u'recovery', u'fever', u'lab', u'versus', u'bread', u'forces',
u'bruce', u'aunt', u'pursuing', u'wk', u'catering', u'applied', u'dumped', u'client', u'elderly',
u'rented', u'medication', u'religiously', u'battle', u'permanently', u'angels', u'cab',
u'economy', u'product', u'produce', u'grandson', u'roommates', u'january', u'transition',
u'drove', u'incarcerated', u'actively', u'fascinated', u'wheels', u'rebuild', u'coworkers', u'taxi',
u'death', u'bryan', u'amy', u'twelve', u'hustle', u'hectic', u'invested', u'restless', u'storage',
u'overseas', u'splitting', u'foster', u'fence', u'heartbroken', u'injury', u'stayed', u'fallen',
u'banks', u'mcdonalds', u'referred', u'surrounded', u'scheduled', u'sisters', u'selling',
u'trapped', u'getter', u'bride', u'salary', u'heavens', u'famous', u'grandparents', u'buisness',
u'coarse', u'partly', u'loop', u'conflict', u'temporary', u'retirement', u'remaining', u'dearly',
u'rut', u'trucking', u'disclosure', u'accustomed', u'practically', u'outing', u'wizard',
u'comforts', u'growing', u'stage', u'mode', u'improving', u'adopted', u'attack', u'final',
u'relatives', u'tire', u'employer', u'driveway', u'nightmare', u'torn', u'bucks', u'paycheck',
u'restore', u'barclub', u'loss', u'payments', u'heartache', u'doctors', u'technician', u'signed',
u'penny', u'contest', u'togather', u'portion', u'sweetest', u'teaching', u'rescue', u'companies',

u'convenience', u'wreck', u'nanny', u'stressing', u'aging', u'president', u'operate',
u'arrested', u'lucy', u'dawn', u'arrived', u'grade', u'dwell', u'villages', u'hadnt', u'anxiety',
u'niece', u'doc', u'folks', u'testing', u'weary', u'modeling', u'sooooo', u'employees',
u'unfortunate', u'claims', u'destroy', u'vanity', u'deaf', u'functions', u'stages', u'sticks',
u'handed', u'dies', u'managed', u'die', u'diagnosed', u'flexibility', u'sake', u'sour', u'kiddos',
u'jail', u'residence', u'police', u'finance', u'killed', u'venture', u'deployed', u'nearly', u'checks',
u'jr', u'cancer', u'loosing', u'speech', u'struggling', u'wives', u'bank', u'careers', u'grid',
u'served', u'pockets', u'mondays', u'dollar', u'shattered', u'mothers', u'lawyer', u'diamond',
u'nineteen', u'manscaping', u'purchased', u'messed', u'pad', u'stepped', u'hart', u'feb',
u'pains', u'horizons', u'outta', u'paths', u'majority', u'eggs', u'expensive', u'gap', u'survivor',
u'itch', u'organization', u'cue', u'magnificent', u'continuing', u'graduation', u'jimmy',
u'notion', u'unsatisfied', u'nephew', u'depression', u'courting', u'breakup', u'waters',
u'chasing', u'retail', u'owning', u'messing', u'completed', u'visited', u'comfortably',
u'payment', u'stranded', u'portfolio', u'climate', u'finished', u'claimed', u'introduced',
u'childrens', u'guard', u'carried', u'fabulous', u'beaten', u'revolves', u'liberty', u'supplies',
u'kicked', u'naive', u'mc', u'my', u'returning', u'toxic', u'destroyed', u'ran', u'celibate', u'bait',
u'avenues', u'swept', u'habit', u'grandmother', u'boss', u'expense', u'breaks', u'burnt',
u'successfully', u'proudly', u'parenting', u'civil', u'solar', u'neighbors', u'row', u'philippines',
u'paralyzed', u'crisis', u'bases', u'employee', u'cases', u'dropped', u'counting', u'joint',
u'furnished', u'factory', u'attended', u'childhood', u'marines', u'pre', u'mates', u'adopt',
u'uncle', u'ssi', u'duty', u'skeletons', u'reminded', u'shifts', u'nest', u'gather', u'depot',
u'quitting', u'royals', u'spends', u'cove', u'exhausting', u'hmmm', u'program', u'siblings',
u'fam', u'grandfather', u'crush', u'version', u'semester', u'nephews', u'remodeling',
u'finances', u'travels', u'stale', u'alike', u'donor', u'seasons', u'fond', u'longest', u'sustainable',
u'defense', u'youth', u'sold', u'betrayed', u'pt', u'brick', u'frozen', u'advocate', u'meds',

u'lone', u'filed', u'junior', u'windy', u'youngest', u'separation', u'fortunate', u'struggled',
u'toddler', u'flex', u'surroundings', u'thousands', u'cried', u'painted', u'crappy', u'barely',
u'member', u'diploma', u'fighting', u'bell', u'empire', u'testosterone', u'transfer', u'funds',
u'boom']

26: CODEWORDS

[u'headline', u'sign', u'favorite', u'subject', u'spam', u'line', u'code', u'zip', u'animal',
u'spammers', u'avoid', u'word', u'purple', u'title', u'fruit', u'put', u'header', u'todays',
u'heading', u'fave', u'folder', u'subj', u'box', u'bot', u'shoe', u'band', u'flavor', u'eye',
u'eliminate', u'filter', u'zodiac', u'season', u'color', u'weed', u'flower', u'song', u'fav']

27: TYPES OF RELATIONSHIPS

[u'weekly', u'hooking', u'flings', u'hookups', u'sex', u'pals', u'occasional',
u'meetings', u'causal', u'platonic', u'pressure', u'dl', u'encounters', u'night', u'anonymous',
u'fling', u'quickie', u'strings', u'random', u'casual', u'fwbs', u'anon', u'bud', u'togethers',
u'buddy', u'third', u'hook', u'regular', u'basis', u'nighters', u'thing', u'sum', u'pal',
u'threesome', u'reciprocation', u'ongoing', u'fwb', u'benefits', u'joining', u'pen', u'nsa', u'reg',
u'sexting', u'rendezvous', u'instant', u'encounter', u'join', u'mm', u'group', u'nite',
u'playtime', u'quick', u'stands', u'nighter', u'buddies', u'stand', u'strictly', u'recip', u'ups',
u'hookup', u'meaningless']

28 : NAMES, TERMS OF ADDRESS

[u'here', u'olds', u'hows', u'rick', u'dave', u'eighteen', u'hay', u'joseph', u'honey', u'jim', u'hey', u'johnny', u'hello', u'adam', u'richard', u'stated', u'scott', u'names', u'horned', u'um', u'babe', u'david', u'jay', u'sweetie', u'btw', u'checking', u'wassup', u'daniel', u'john', u'ww', u'brad', u'tony', u'josh', u'ole', u'angel', u'mrs', u'soldier', u'robert', u'jeff', u'anthony', u'michael', u'ryan', u'kenny', u'ben', u'ashley', u'hi', u'eric', u'fellas', u'don', u'm', u'sam', u'whats', u'heres', u'boy', u'boo', u'dude', u'max', u'tyler', u'kevin', u'sean', u'mark', u'alex', u'chris', u'everyone', u'thomas', u'hey', u'steven', u'jason', u'named', u'dan', u'matt', u'jessica', u'bye', u'chick', u'justin', u'j', u'mw', u'mr', u'ima', u'howdy', u'ray', u'brandon', u'stopping', u'tim', u'ron', u'rob', u'yall', u'says', u'lady', u'andrew', u'ol', u'ladies', u'gary', u'taylor', u'sup', u'gals', u'prince', u'nick', u'steve', u'brian', u'joe', u'mike', u'name', u'james']

29: NOT CONSISTENT

[u'txt', u'fiv', u'won', u'ate', u'tree', u'sevento', u'one', u'ty', u'sevenfive', u'ur', u'sevenseven', u'sevenzero', u'l', u'intrested', u'com', u'nin', u'thx', u'thanks', u'shout', u'pls', u'plz', u'p', u'yeah', u'asap', u'soon', u'yu', u'svn', u'smarty', u'ure', u'fiveseven', u'call', u'dot', u'yea', u'fore', u'holla', u'thre', u'thanx', u'\U0001f609', u'\U0001f60a', u'xoxo', u'picpic', u'u', u'please', u'lt', u'sevenfiveseven', u'hear', u'pixs', u'msg', u'fourty', u'seventoo', u'r', u'oo', u'oh', u'tex', u'waiting', u'genius', u'seventwo', u'hmu', u'loney']

30: NOT CONSISTENT

[u'prices', u'prize', u'elaborate', u'explained', u'replace', u'spoke', u'catchy', u'passport', u'strike', u'relay', u'example', u'caution', u'complaining', u'hoo', u'machine', u'classify', u'typed', u're', u'selfs', u'wast', u'tango', u'showed', u'dozen', u'thrilled', u'landmark', u'entry', u'net', u'pressured', u'wth', u'pity', u'reported', u'doubts', u'qualifications', u'browse', u'sob', u'proves', u'solicitation', u'holy', u'choice', u'prevent', u'tasks', u'introducing', u'easiest', u'deserved', u'muah', u'instantly', u'compelled', u'entering', u'internet', u'initiate', u'indicate', u'argument', u'foward', u'awaiting', u'borrow', u'screen', u'textemail', u'thousand', u'verified', u'cheers', u'haystack', u'vague', u'stranger', u'gimme', u'scamming', u'refer', u'texted', u'original', u'returned', u'method', u'revealing', u'negativity', u'paper', u'signs', u'bypass', u'weeds', u'reduce', u'research', u'qualify', u'anonymity', u'sell', u'incase', u'cover', u'insane', u'hint', u'viewing', u'input', u'defiantly', u'emergency', u'conversing', u'bk', u'by', u'garbage', u'hometown', u'elsewhere', u'poster', u'fades', u'uh', u'ud', u'essay', u'results', u'categories', u'timely', u'consistently', u'purposes', u'delay', u'reposting', u'await', u'embarrass', u'spill', u'false', u'assholes', u'pause', u'following', u'invited', u'drank', u'hateful', u'hahaha', u'lest', u'wen', u'applicants', u'fools', u'choosing', u'deciding', u'letters', u'mere', u'hehe', u'shyness', u'official', u'reciprocated', u'nonetheless', u'rule', u'compete', u'phones', u'approached', u>equals', u'items', u'insult', u'tolerated', u'tinder', u'disappear', u'talktext', u'cuties', u'meets', u'nigerian', u'standard', u'organize', u'latter', u'explanation', u'weeding', u'reserve', u'besties', u'cow', u'brag', u'publicly', u'applies', u'cons', u'jerry', u'advertising', u'the', u'repost', u'wright', u'boxes', u'peek', u'pose', u'trophy', u'allways', u'accordingly', u'prayer', u'test', u'unwilling', u'compensate', u'brownie', u'responders', u'curse', u'hoops', u'trial', u'marked', u'dateing', u'topic', u'exclude', u'explicit', u'correspondence', u'lmk',

u'unwanted', u'inevitable', u'invitation', u'blind', u'checked', u'craig', u'online', u'price',
u'cannot', u'countless', u'trick', u'figuring', u'closest', u'struck', u'politely', u'miracle',
u'chop', u'prepare', u'bucket', u'valid', u'anybodys', u'selfies', u'inspired', u'phrase', u'reject',
u'identity', u'nigeria', u'boredom', u'useless', u'cavs', u'quote', u'usual', u'okcupid',
u'descriptions', u'woo', u'distinguish', u'phonies', u'suggestion', u'dime', u'chore', u'chord',
u'resorted', u'block', u'placing', u'properly', u'application', u'refrain', u'ignorant', u'gh',
u'maam', u'yourselves', u'download', u'pitch', u'camera', u'spelling', u'agenda', u'crossing',
u'route', u'instance', u'legitimate', u'scrolling', u'correctly', u'yay', u'yah', u'piques', u'novel',
u'xoxoxo', u'generic', u'discouraged', u'vids', u'twitter', u'reaching', u'hv', u'neutral',
u'reservations', u'preggo', u'talkin', u'updated', u'million', u'drew', u'intrigue', u'whatnot',
u'recommend', u'warn', u'setup', u'untill', u'thankyou', u'purchase', u'attempt', u'goodbye',
u'wud', u'gunna', u'resume', u'messaged', u'guidelines', u'fitting', u'proving', u'attempts',
u'comply', u'bam', u'reference', u'ignorance', u'damned', u'ik', u'in', u'mouse', u'picts',
u'browsing', u'facts', u'save', u'whichever', u'paragraphs', u'dear', u'arguing', u'nerve',
u'selfie', u'pages', u'failure', u'skip', u'disgusting', u'marks', u'accidentally', u'attracting',
u'\U0001f48b', u'useful', u'soup', u'ull', u'sample', u'listing', u'emailed', u'\U0001f618',
u'\U0001f601', u'\U0001f600', u'\U0001f60d', u'correct', u'continuous', u'occupied',
u'wasters', u'shocked', u'insanity', u'complaints', u'trolling', u'traded', u'comprehend',
u'attaching', u'promising', u'hmmmm', u'bags', u'pan', u'disappointment', u'weave',
u'amazed', u'images', u'references', u'arranged', u'file', u'bragging', u'psychic', u'reluctant',
u'celebrity', u'titles', u'score', u'unappreciated', u'crowded', u'possessions', u'thnx',
u'acquaintances', u'print', u'lease', u'trolls', u'mahalo', u'ruin', u'idiots', u'permission',
u'donate', u'floats', u'oovoo', u'chatted', u'task', u'alternate', u'bin', u'wil', u'rebound',
u'judgement', u'google', u'examples', u'contacts', u'use', u'lottery', u'goods', u'delusional',
u'hoped', u'rejection', u'theyll', u'statement', u'anonymously', u'explaining', u'groceries',

u'objects', u'limb', u'bothered', u'dam', u'spell', u'lecture', u'clarify', u'monkey', u'selection',
u'interview', u'attempting', u'glorious', u'comment', u'relevant', u'fucker', u'judgment',
u'wether', u'noise', u'\uf04a', u'ramble', u'yell', u'ttyl', u'ttys', u'regrets', u'venue',
u'hundreds', u'digits', u'forum', u'stray', u'befor', u'confuse', u'signing', u'adjust', u'titled',
u'ton', u'murder', u'inspire', u'contacted', u'bail', u'report', u'automatic', u'confusion',
u'participating', u'spambots', u'suggest', u'characters', u'shortly', u'laptop', u'whomever',
u'advertise', u'discount', u'accommodate', u'rounds', u'rely', u'discourage', u'removed',
u'willingly', u'papers', u'impression', u'cyber', u'iphone', u'avengers', u'qualified', u'insist',
u'hun', u'hurtful', u'cops', u'copy', u'ans', u'donation', u'cheap', u'apologies', u'reaction',
u'contract', u'pof', u'tittle', u'cast', u'check', u'propose', u'disrupt', u'donations', u'gmail',
u'poem', u'update', u'creepers', u'ot', u'accounts', u'lastly', u'applying', u'beds', u'idiot',
u'farther', u'briefly', u'proposition', u'abit', u'dish', u'everyones', u'list', u'hesitation', u'flaw',
u'court', u'desperation', u'motels', u'explains', u'reflect', u'shady', u'prolly', u'convinced',
u'instruction', u'stall', u'dang', u'entirety', u'unrealistic', u'blank', u'guts', u'pp', u'rid',
u'lengthy', u'fate', u'negotiate', u'stating', u'unlimited', u'collect', u'ticking', u'prob',
u'impressed', u'motive', u'warning', u'hooks', u'disclaimer', u'pleas', u'goodies', u'ummm',
u'suggests', u'slightest', u'fee', u'tab', u'instead', u'somethin', u'crash', u'edit', u'disclose',
u'unknown', u'tommy', u'nobodys', u'candidates', u'words', u'exam', u'hmm', u'spammed',
u'disagree', u'romeo', u'approve', u'obtain', u'leap', u'locate', u'mite', u'owe', u'appt', u'apps',
u'temp', u'arnt', u'demonstrate', u'cmon', u'junk']

31: OUTDOOR LEISURE ACTIVITIES

[u'woods', u'bbqs', u'cooking', u'travel', u'camp', u'adventures', u'canoeing', u'skiing', u'cookouts', u'bowling', u'zoo', u'grilling', u'trips', u'vacations', u'indoors', u'mall', u'golf', u'river', u'spending', u'casino', u'mudding', u'riding', u'bike', u'casinos', u'hike', u'amusement', u'horseback', u'hiking', u'concerts', u'fires', u'theater', u'swim', u'fairs', u'dancing', u'pool', u'kayaking', u'trails', u'dinning', u'biking', u'swimming', u'staying', u'baking', u'movies', u'boating', u'chilling', u'sporting', u'bon', u'beach', u'mountains', u'camping', u'shopping', u'outdoor', u'karaoke', u'motorcycles', u'restaurants', u'gardening', u'sunsets', u'festivals', u'dinners', u'parks', u'walking', u'walks', u'lake', u'hikes', u'enjoy', u'hunting', u'museums', u'rides', u'ocean', u'getaways', u'bonfires', u'picnics', u'dining', u'events', u'beaches', u'fish', u'activities', u'markets', u'outdoors', u'park', u'wheeling', u'traveling', u'flea', u'watching', u'bbq', u'fishing', u'wheelers']

32: POSITIVE PHYSICAL EXPERIENCE

[u'relieving', u'release', u'desire', u'offer', u'erotic', u'stress', u'rubdown', u'experience', u'session', u'pleasures', u'healing', u'enjoyable', u'engage', u'soothing', u'lust', u'moments', u'passion', u'excitement', u'safe', u'seduction', u'techniques', u'warmth', u'tissue', u'orgasms', u'stimulating', u'dominance', u'enjoyment', u'orgasm', u'sensual', u'massage', u'submission', u'therapeutic', u'smiles', u'pure', u'powerful', u'providing', u'tantra', u'skills', u'crave', u'pleasurable', u'combination', u'ending', u'offering', u'sensuous', u'relieve', u'satisfying', u'affection', u'deep', u'lovemaking', u'pleasure', u'energy', u'engaging', u'tender', u'pleasuring', u'exploration', u'oils', u'fulfillment', u'appetite', u'pampering', u'intimacy', u'environment', u'sessions', u'laughter', u'swedish', u'uninhibited', u'stimulation', u'satisfaction', u'longing', u'relief, u'womans', u'tantric',

u'relaxation', u'closeness', u'consensual', u'attention', u'intense', u'fulfilling', u'tenderness',
u'desired', u'tension', u'incredible']

33: UNDESIRABLE CHARACTERISTICS & BEHAVIOR

[u'immature', u'tired', u'weirdos', u'men', u'haters', u'threesomes', u'fakers', u'fake',
u'allowed', u'prostitutes', u'disrespect', u'crazies', u'worries', u'liers', u'escorts', u'sums',
u'period', u'hookers', u'scam', u'bullshitters', u'trannies', u'inquires', u'singles', u'scams',
u'bs', u'robots', u'scammer', u'gays', u'waste', u'site', u'web', u'cheats', u'fakes', u'boyfriends',
u'comments', u'wasting', u'card', u'cheaters', u'losers', u'creeps', u'creepy', u'chicks',
u'pretending', u'bisexuals', u'bots', u'nonsense', u'crap', u'attachments', u'bullshit', u'fats',
u'credit', u'diggers', u'website', u'lies', u'shit', u'pervs', u'guys', u'stupid', u'queens', u'sites',
u'those', u'profiles', u'harm', u'childish', u'jerks', u'whores', u'somes', u'bis', u'collectors',
u'excuses', u'bunch', u'bull', u'takers', u'ads', u'flake', u'whatsoever', u'replys', u'phony',
u'players', u'games', u'no', u'posts', u'repeat', u'offers', u'ppl', u'replies', u'couple', u'catfish',
u'flakes', u'trash', u'websites', u'liars', u'perverts', u'inquiries', u'dudes', u'ps', u'pros',
u'apply', u'spams', u'males', u'scammers', u'responses', u'craigs', u'intention', u'sick']

34: PHOTOS & REQUIRED INFORMATION

[u'clothed', u'assured', u'needed', u'current', u'photos', u'agestats', u'torso',
u'detailed', u'quicker', u'pict', u'information', u'introduction', u'immediate', u'occupation',
u'receive', u'recent', u'picks', u'location', u'send', u'sent', u'consideration', u'gladly',

u'inappropriate', u'face', u'upon', u'details', u'upload', u'provide', u'rated', u'rates',
u'sentence', u'ill', u'sends', u'replay', u'initial', u'yourself', u'appreciated', u'inquiry',
u'picsstats', u'responce', u'your', u'sentences', u'clear', u'x', u'info', u'required', u'fastest',
u'attach', u'photograph', u'urs', u'contain', u'stats', u'ours', u'accurate', u'immediately',
u'necessary', u'additional', u'reciprocate', u'pg', u'stat', u'intrest', u'collector', u'mine',
u'basic', u'ig', u'yours', u'nude', u'automatically', u'dic', u'summary', u'nudes', u'urself',
u'confirm', u'inbox', u'minimum', u'response', u'services', u'ph', u'bio', u'requesting',
u'liners', u'prepared', u'lines', u'liner', u'descriptive', u'picno', u'enclose', u'otherwise',
u'discretion', u'statspics', u'pic', u'pix', u'detail', u'picstats', u'request', u'description',
u'availability', u'photo', u'sunglasses', u'shots', u'recieve', u'num', u'expected', u'deleted',
u'statspic', u'discription', u'picture', u'tasteful', u'facebody', u'paragraph', u'reply', u'yur',
u'favor', u'ignored', u'telephone', u'without', u'provided', u'shirtless', u'intro', u'requested',
u'included', u'rate', u'brief', u'directions', u'return', u'include', u'proof', u'promptly', u'g',
u'will', u'supply', u'gets', u'helpful', u'mines"]

35: TIMES&WEATHER

[u'rainning', u'wednesday', u'fri', u'morning', u'sat', u'vacation', u'summer', u'th',
u'week', u'tonite', u'sunny', u'tonight', u'monday', u'thursday', u'memorial', u'sun',
u'available', u'tomorrow', u'weekdays', u'evenings', u'anytime', u'evening', u'holiday',
u'during', u'afternoons', u'tuesday', u'hosting', u'daytime', u'afternoon', u'weekend',
u'sunday', u'town', u'late', u'mornings', u'noon', u'weekday', u'day', u'early', u'weekends',
u'saturday', u'motel', u'sometime', u'cloudy', u'today', u'birthday', u'midnight', u'mon',
u'friday', u'hotel', u'nights', u'pm', u'rainy', u'july', u'weather', u'june"]

36: NOT CONSISTENT

[u'relax', u'hold', u'listen', u'doors', u'chill', u'jerk', u'flirt', u'laying', u'play', u'cry', u'kiss', u'cuddle', u'lay', u'hang', u'whenever', u'flowers', u'asleep', u'snuggle', u'sneak', u'fool', u'while', u'bed', u'kick', u'eat', u'stay', u'wake', u'fall', u'hangout', u'couch', u'curl', u'vent', u'sleep', u'jack', u'gentlman', u'mess', u'watch', u'hug', u'cuddled', u'come', u'sit']

37: SPANISH WORDS

[u'hola', u'rico', u'k', u'fotos', u'buscando', u'o', u'soy', u'ver', u'su', u'si', u'se', u'al', u'tu', u'te', u'ser', u'sea', u'un', u'h', u'de', u'bien', u'con', u'el', u'en', u'es', u'buen', u'tengo', u'cuerpo', u'hablo', u'todo', u'pero', u'foto', u'una', u'mujer', u'hombre', u'q', u'estoy', u'para', u'busco', u'y', u'le', u'la', u'lo', u'que', u'tiempo', u'tambien', u'quiero', u'mi', u'muy', u'ni', u'espanol', u'gusta', u'como', u'por', u'mas', u'solo', u'sin']

38: SEX TERMS& TERMS OF DOMINATIO

[u'cocksucker', u'orally', u'service', u'master', u'obedient', u'birth', u'verbally', u'cherished', u'sexually', u'submit', u'mans', u'abused', u'mistress', u'pleasured', u'spoiled', u'eager', u'deserves', u'touched', u'fulfill', u'owned', u'command', u'total', u'satisfied', u'perform', u'cherish', u'satisfy', u'pillow', u'serving', u'servant', u'shown', u'breed',

u'daddy', u'cared', u'respected', u'touch', u'versa', u'fully', u'dominated', u'trained', u'taken',
u'daddys', u'held', u'desperately', u'pamper', u'alpha', u'dominant', u'guide', u'worshipped',
u'humiliated', u'control', u'princess', u'bitch', u'romanced', u'punish', u'adored', u'sissy',
u'taught', u'surprise', u'naughty', u'disciplined', u'sees', u'dom', u'unconditionally',
u'surrender', u'switch', u'maid', u'dominate', u'protected', u'spank', u'fantasy', u'needs',
u'treated', u'orders', u'serviced', u'craves', u'discipline', u'pet', u'cater', u'queen', u'train',
u'protect', u'experienced', u'loved', u'pampered', u'position', u'proud', u'ropes', u'domme',
u'secretly', u'toy', u'serve', u'treats', u'goddess', u'submissive', u'humiliate', u'adore',
u'charge', u'advantage', u'pleased', u'worship', u'strict', u'treat', u'controlled', u'deserve',
u'treating', u'sub', u'teach', u'king', u'him', u'wishes', u'sir', u'supports', u'spoil', u'whore',
u'slave', u'obey', u'slut']

39 AGE & ETHNICITY

[u'forties', u'preferable', u'vgl', u'lookin', u'femm', u'eastern', u'european', u'old',
u'filipino', u'filipina', u'topvers', u'spanish', u'twenties', u'hawaiian', u'aggressive', u's',
u'jamaican', u'am', u'masc', u'loooking', u'mexican', u'w', u'aaf', u'straight', u'couple',
u'curious', u'bm', u'jewish', u'young', u'twink', u'indian', u'puerto', u'latina', u'korean',
u'biker', u'btm', u'inexperienced', u'yrs', u'bi', u'tops', u'thirties', u'wf', u'wm', u'descent',
u'fitathletic', u'arab', u'latino', u'year', u'professionals', u'rugged', u'prefers', u'feminine',
u'professional', u'unmarried', u'gwm', u'gentlemen', u'verse', u'dominican', u'lipstick',
u'singledivorced', u'vers', u'yo', u'yr', u'american', u'str', u'stem', u'gl', u'gent', u'fella',
u'businessman', u'femme', u'asian', u'lesbian', u'biracial', u'iso', u'hottie', u'closeted',
u'visitor', u'iam', u'chub', u'bicurious', u'ssbbw', u'inshape', u'milf', u'preferrably', u'boi',

u'african', u'cute', u'black', u'mixed', u'straightbi', u'search', u'rican', u'seeking', u'acting',
u'sbf', u'sbm', u'caucasion', u'biwm', u'aged', u'gay', u'aa', u'lds', u'cub', u'masculine', u'f',
u'native', u'frat', u'dad', u'stud', u'chaser', u'jocks', u'bisexual', u'mid', u'mix', u'fellow',
u'citizen', u'female', u'versatile', u'mw'f', u'mwm', u'seeks', u'ebony', u'newly', u'hispanic',
u'queer', u'attractive', u'crossdresser', u'sixties', u'farmer', u'white', u'senior', u'fifties',
u'hisp', u'poz', u'oriental', u'caucasian', u'exec', u'latin', u'swm', u'sw'f', u'prof', u'youthful',
u'sexy', u'trans', u'fem', u'dw'f', u'dw'm', u'redhead', u'bttm', u'wht', u'bbw', u'blk',
u'attractive', u'goodlooking']

40: MARRIAGE & CHILDREN

[u'kids', u'cheating', u'partnered', u'babies', u'son', u'boyfriend', u'engaged',
u'hubby', u'widowed', u'hes', u'teen', u'male', u'neglected', u'marriage', u'seperated',
u'parents', u'wives', u'bf', u'marry', u'involved', u'baby', u'sister', u'child', u'daughter',
u'separated', u'divorced', u'already', u'roommate', u'grandchildren', u'happily',
u'unhappily', u'brothers', u'ex', u'unemployed', u'husband', u'family', u'broke', u'fiance',
u'step', u'gf', u'widow', u'sole', u'he', u'lives', u'grandkids', u'grown', u'kid', u'yes',
u'attached', u'father', u'teenage', u'man', u'parent', u'custody', u'mom', u'girlfriend',
u'children', u'biological', u'two', u'sexless', u'pregnant', u'mama', u'unhappy', u'raise',
u'husbands', u'boys', u'single', u'widower', u'loveless', u'mother', u'dads', u'teenager', u'hsv',
u'herpes', u'moms', u'legally', u'ok', u'spouse', u'wife', u'sons', u'teens', u'brother', u'divorce',
u'exes', u'daughters', u'mommy', u'momma', u'girlfriends', u'raising', u'someday',
u'teenagers', u'married']

41: ??

[u'females', u'preferably', u'types', u'timers', u'welcomed', u'fems', u'twinks', u'prefer', u'offense', u'preference', u'racial', u'stone', u'consider', u'factor', u'races', u'religion', u'appropriate', u'hangups', u'age', u'legal', u'race', u'agerace', u'ethnicities', u'negotiable', u'chubs', u'above', u'studs', u'welcome', u'typically', u'between', u'stems', u'daddies', u'issue', u'tomboys', u'older', u'idc', u'dosent', u'bears', u'femmes', u'bottoms', u'sizes', u'preferences', u'exceptions', u'marital', u'girls', u'matter', u'shapes', u'discriminate', u'acceptable', u'unimportant', u'label', u'nationality', u'whites', u'difference', u'gender', u'ages', u'butches', u'matters', u'preferred', u'under', u'colors', u'asians', u'latinas', u'younger', u'groups', u'pref, u'requirements', u'attracted', u'encouraged', u'irrelevant', u'status', u'perfer', u'blacks', u'lesbians', u'cds', u'ethnicity', u'range', u'women', u'ethnic', u'latinos', u'bbws', u'closer', u'limit']

42: BODY SIZE & WEIGHT

[u'fit', u'obese', u'slightly', u'chubby', u'average', u'slender', u'large', u'small', u'shorter', u'thin', u'slim', u'frame', u'figured', u'voluptuous', u'smooth', u'curvy', u'thick', u'extra', u'size', u'chunky', u'heavy', u'uncut', u'overweight', u'butch', u'taller', u'sized', u'ish', u'girly', u'height', u'thinner', u'carry', u'heavier', u'husky', u'petite', u'shape', u'big', u'proportional', u'bones', u'curves', u'skinny', u'thicker', u'proportioned', u'athletic', u'hourglass', u'hairy', u'plus', u'busty', u'decent', u'stocky', u'fat', u'short', u'bigger', u'weight', u'tomboy', u'smaller', u'larger', u'avg']

43: PREFERENCES IN SPORTS & MUSIC & FOOD

[u'specially', u'hero', u'golden', u'foodie', u'dragons', u'yum', u'arrow', u'snakes', u'series', u'musicians', u'ranging', u'trek', u'singer', u'marvel', u'tequila', u'menu', u'pocket', u'rings', u'fruits', u'electro', u'survival', u'sox', u'tigers', u'seafood', u'actor', u'reggae', u'chain', u'osu', u'skate', u'cooks', u'team', u'such', u'quarter', u'poems', u'chevy', u'potter', u'wandering', u'warriors', u'backpack', u'skies', u'tube', u'tropical', u'crowds', u'watcher', u'tail', u'cable', u'harry', u'ridding', u'plant', u'author', u'astrology', u'sauce', u'fights', u'coins', u'pickup', u'packers', u'ducks', u'cuisine', u'salutations', u'wildlife', u'repair', u'arcade', u'custom', u'camo', u'walker', u'gossip', u'magic', u'champagne', u'twins', u'bird', u'tools', u'charity', u'rods', u'npr', u'varies', u'instruments', u'cuss', u'touring', u'ordering', u'collecting', u'dishes', u'soap', u'mechanical', u'jam', u'bourbon', u'superhero', u'etc', u'laser', u'ing', u'ghost', u'porno', u'horn', u'study', u'era', u'production', u'network', u'medicine', u'recovering', u'electronic', u'burgers', u'circles', u'dj', u'leisure', u'dr', u'cattle', u'discussions', u'tons', u'duties', u'vegetarian', u'caps', u'marketing', u'war', u'wwe', u'gun', u'v', u'snowboard', u'spontaneity', u'graphic', u'magazine', u'birds', u'teams', u'album', u'british', u'nutrition', u'quad', u'thai', u'equipment', u'sport', u'media', u'favorites', u'livestock', u'greetings', u'improvement', u'ie', u'log', u'housework', u'dabble', u'vs', u'cookies', u'tracks', u'cows', u'systems', u'fees', u'houses', u'archery', u'physics', u'coaster', u'gloves', u'league', u'abroad', u'specialty', u'electrical', u'motorhome', u'suits', u'wave', u'commercial', u'products', u'halloween', u'bathing', u'including', u'auto', u'entertainment', u'wines', u'gallery', u'landscaping', u'tunes', u'computer', u'supernatural', u'fishin', u'paranormal', u'puppies', u'pc', u'hd', u'mechanics', u'universal', u'consists', u'captain',

u'japanese', u'guests', u'vodka', u'programming', u'therapy', u'acoustic', u'modern', u'mint',
u'davidson', u'mario', u'points', u'lyrics', u'fields', u'warrior', u'locations', u'cane',
u'renaissance', u'programs', u'background', u'performing', u'zombies', u'batman', u'ford',
u'digital', u'utilities', u'diet', u'reptiles', u'tacos', u'sappy', u'international', u'stormy', u'bow',
u'bob', u'sword', u'classics', u'settings', u'conventions', u'cheesy', u'specials', u'gathering',
u'performance', u'channel', u'enthusiast', u'dice', u'spirituality', u'seas', u'trades', u'mary',
u'extensive', u'lightning', u'fanatic', u'videos', u'deer', u'weights', u'ranges', u'bible',
u'published', u'industrial', u'mystery', u'disneyland', u'\u2022', u'gospel', u'repairs', u'fans',
u'myriad', u'iron', u'foreign', u'gatherings', u'italy', u'development', u'subjects', u'jewelry',
u'aliens', u'favourite', u'recreation', u'stock', u'philosophy', u'collection', u'labor', u'junkie',
u'electronics', u'math', u'ancient', u'liquor', u'gear', u'halo', u'muddy', u'latest', u'seats',
u'articles', u'esp', u'shelter', u'chains', u'vegetables', u'pursuits', u'astronomy', u'baker',
u'estate', u'salsa', u'plumbing', u'tourist', u'potatoes', u'fitness', u'automotive', u'urban',
u'mixture', u'countries', u'discovery', u'news', u'towels', u'bees', u'specialize', u'corn',
u'yacht', u'discovering', u'stream', u'sailboat', u'ballroom', u'obsession', u'profit', u'theory',
u'array', u'pokemon', u'flicks', u'dramas', u'crystal', u'engine', u'cash', u'carpentry', u'derby',
u'persona', u'cartoon', u'styles', u'seahawks', u'furniture', u'cowboys', u'amongst',
u'chinese', u'meditation', u'design', u'nba', u'mma', u'minor', u'pony', u'pond', u'lighting',
u'hunter', u'sailor', u'gourmet', u'electric', u'stones', u'sleepovers', u'barn', u'cosplay',
u'various', u'instructor', u'plants', u'homework', u'chickens', u'dvds', u'zombie',
u'housekeeping', u'consist', u'z', u'edm', u'languages', u'deals', u'apple', u'duck', u'eagles',
u'skydiving', u'cardio', u'martini', u'farming', u'billy', u'trap', u'paddling', u'artists',
u'related', u'organic', u'sew', u'shelf', u'combat', u'debate', u'dope', u'chips', u'beverages',
u'gothic', u'vintage', u'celtic', u'console', u'wilderness', u'van', u'book"]

44: NUMBERS

[u'four', u'inches', u'eleven', u'five', u'forty', u'fifty', u'sixty', u'nine', u'foot', u'hundred', u'ninety', u'seventy', u'three', u'seven', u'zero', u'twenty', u'eight', u'eighty', u'ten', u'thirty', u'six']

45: RELATIONSHIPS

[u'foundation', u'previous', u'possible', u'forming', u'eventual', u'interracial', u'establishing', u'possibility', u'future', u'companionship', u'situation', u'intimate', u'relationship', u'dating', u'cuckold', u'mutually', u'partnership', u'romance', u'commitment', u'relationships', u'meaningful', u'lifelong', u'ds', u'scenario', u'beneficial', u'becoming', u>true', u'triad', u'form', u'developing', u'equal', u'develop', u'benefit', u'building', u'committed', u'actual', u'arrangement', u'affair', u'partners', u'rewarding', u'dynamic', u'distance', u'potentially', u'commit', u'ship', u'grow', u'potential', u'lifetime', u'establish', u'lovers', u'secret', u'solid', u'hopes', u'friendshiprelationship', u'sexual', u'sdsb', u'committed', u'friendships', u'monogamous', u'lasting', u'longterm', u'relations', u'exclusive', u'term', u'ultimately', u'anr', u'emphasis', u'connections', u'poly', u'domsub', u'friendship', u'leading', u'option', u'traditional', u'polyamorous', u'activity', u'bond', u'relation', u'permanent', u'ltr']

46: NOT CONSISTENT

[u'therefore', u'rich', u'mandatory', u'creep', u'opposed', u'slob', u'worried', u'pushy', u'rush', u'picky', u'into', u'definitely', u'concerned', u'thats', u'perfect', u'anyones', u'means', u'does', u'problem', u'hooker', u'steal', u'tho', u'cheater', u'certainly', u'guy', u'barbie', u'super', u'offend', u'typical', u'not', u'nor', u'wont', u'care', u'aint', u'concern', u'worry', u'poor', u'unattractive', u'religious', u'arent', u'offended', u'rude', u'materialistic', u'expecting', u'lie', u'cocky', u'wouldnt', u'pervert', u'jump', u'supermodel', u'desperate', u'neither', u'ken', u'digger', u'dose', u'sorry', u'complicated', u'type', u'mean', u'mind', u'sugar', u'judged', u'lazy', u'upfront', u'against', u'ugly', u'jealous', u'deal', u'clingy', u'judging', u'doll', u'gross', u'overly', u'liar', u'particular', u'needy', u'either', u'arrogant', u'disrespectful', u'dont', u'particularly', u'fine', u'superficial', u'ashamed', u'serial', u'perfection', u'material', u'although', u'doesnt', u'prostitute', u'labels', u'however', u'judge', u'specific', u'standards', u'prude', u'necessarily', u'any', u'cheat', u'model', u'racist', u'weirdo', u'shouldnt', u'isnt', u'looks', u'shallow', u'afraid', u'requirement', u'breaker', u'morbidly', u'selfish']

47: CONNECTING & 'THE SPARK'

[u'turn', u'foremost', u'rushing', u'possibly', u'begin', u'later', u'date', u'attraction', u'slow', u'where', u'spark', u'evolves', u'see', u'connection', u'continue', u'could', u'move', u'become', u'eventually', u'depends', u'we', u'comfortable', u'turns', u'happen', u'comes', u'start', u'decide', u'develops', u'discussed', u'blossom', u'feels', u'click', u'vibe', u'eachother', u'compatible', u'friends', u'deeper', u'it', u'grows', u'evolve', u'maybe', u'mesh', u'behind', u'acquainted', u'each', u'further', u'public', u'connect', u'chemistry', u'flow', u'becomes',

u'arrange', u'sparks', u'progress', u'agree', u'happens', u'expectations', u'hopefully',
u'serious', u'leads', u'goes', u'possibilities', u'determine', u'depending', u'proceed', u'relate',
u'theres', u'develope', u'itll', u'common', u'agreed', u'closed', u'perhaps', u'discuss', u'lead',
u'clicks']

48: POSITIVE CHARACTERISTICS & ADJECTIVES

[u'looking', u'similarly', u'best', u'fantastic', u'ideal', u'right', u'for', u'happy',
u'settle', u'freaky', u'awesome', u'wanting', u'lonley', u'enjoys', u'wants', u'zest', u'with',
u'ideally', u'company', u'beautiful', u'misses', u'seaching', u'adventurers', u'soulmate',
u'sweetheart', u'real', u'hoping', u'specifically', u'having', u'freind', u'great', u'desiring',
u'cougar', u'accompany', u'lucky', u'kindred', u'ordinary', u'sumone', u'another', u'has',
u'mingle', u'accepts', u'bedroom', u'shares', u'sincerely', u'chic', u'adventure', u'friend',
u'similiar', u'forever', u'ladie', u'somebody', u'lady', u'exciting', u'dream', u'cutie',
u'discreetly', u'wonderful', u'singledivorcedwidowed', u'close', u'missing', u'likes', u'some',
u'similar', u'experiment', u'housewife', u'counterpart', u'quality', u'need', u'equally',
u'whos', u'whose', u'adult', u'amazing', u'girlwoman', u'bestfriend', u'somone', u'lovely',
u'is', u'im', u'discreet', u'befriend', u'wishing', u'girl', u'normal', u'same', u'companion',
u'spice', u'bestie', u'friendlover', u'share', u'good', u'extraordinary', u'needing', u'find',
u'simple', u'gal', u'fun', u'spend', u'truely', u'playmate', u'mate', u'loves', u'lover', u'partner',
u'special', u'consistent', u'mentor', u'cool', u'crime', u'someone', u'confidant', u'woman',
u'hot', u'knows', u'basically', u'benifits', u'finding', u'bfff', u'starters', u'finds', u'whom',
u'whod', u'lovin', u'long', u'explore', u'figure', u'gurl', u'lookn', u'who', u'refreshing',
u'mainly', u'rite', u'searching']

49: INCONSISTENT

[u'lord', u'polyamory', u'pulse', u'consenting', u'psychological', u'stern', u'sweeter', u'mindedness', u'effective', u'daddydaughter', u'wins', u'attracts', u'feedback', u'rewarded', u'effects', u'adapt', u'rewards', u'endure', u'spectrum', u'affairs', u'causing', u'object', u'participate', u'lessons', u'fair', u'playfulness', u'extend', u'extent', u'weirdness', u'damaged', u'harmony', u'ownership', u'tune', u'beings', u'understandable', u'proven', u'intent', u'icing', u'cleanliness', u'brilliant', u'visual', u'valued', u'believed', u'listens', u'explored', u'tendency', u'stupidity', u'matches', u'maintaining', u'canvas', u'intensely', u'training', u'emotion', u'kinkier', u'hedonistic', u'opposite', u'imagined', u'outlet', u'vision', u'expression', u'combined', u'motives', u'influence', u'intimately', u'wired', u'instrument', u'leadership', u'image', u'gifts', u'caused', u'causes', u'norm', u'investment', u'sweetness', u'sexiness', u'keeping', u'resources', u'evolved', u'creature', u'ruining', u'yearns', u'definition', u'servitude', u'internal', u'virus', u'impact', u'writes', u'dependent', u'distraction', u'creator', u'acceptance', u'longs', u'mindset', u'handling', u'forgetting', u'necessity', u'surprises', u'demands', u'quest', u'communications', u'disorder', u'methods', u'behave', u'masculinity', u'agreement', u'span', u'competition', u'likewise', u'defines', u'elements', u'sided', u'existing', u'dynamics', u'expressed', u'led', u'involves', u'subdom', u'clients', u'element', u'realizes', u'insight', u'delightful', u'interactions', u'godly', u'fascinating', u'have', u'courtesy', u'stronger', u'atmosphere', u'familiar', u'destiny', u'motto', u'homosexual', u'greater', u'gaining', u'lifestyle', u'frustration', u'function', u'society', u'contribute', u'compared', u'illusion', u'inspiration', u'protection', u'romantically', u'striving', u'elusive', u'fear', u'jesus', u'backgrounds', u'unfulfilled',

u'creates', u'guilt', u'solely', u'failing', u'respecting', u'manner', u'realm', u'involving',
u'religions', u'respectfully', u'subtle', u'consciousness', u'guidance', u'carries', u'happiest',
u'drawn', u'terms', u'essentially', u'gratitude', u'concerning', u'aid', u'gods', u'honored',
u'hassle', u'confusing', u'taker', u'quirks', u'mysterious', u'physicality', u'maximum',
u'generosity', u'nudity', u'promised', u'mold', u'curiosities', u'believe', u'collared', u'concept',
u'supported', u'presented', u'fragile', u'assistance', u'regardless', u'attribute', u'occur',
u'discussion', u'sacrifice', u'brain', u'acknowledge', u'sexiest', u'forms', u'courtship',
u'thankful', u'obligation', u'tries', u'striking', u'optimist', u'anger', u'objective',
u'attractiveness', u'slaves', u'essence', u'complications', u'informed', u'freely', u'orientation',
u'tradition', u'severe', u'treatment', u'fortune', u'greatly', u'involvement', u'receptive',
u'endeavors', u'confidential', u'amount', u'helps', u'conventional', u'entails', u'lacks',
u'communicating', u'ego', u'solving', u'fairy', u'beyond', u'dominating', u'obedience',
u'toward', u'substantial', u'vulnerable', u'caliber', u'circle', u'protecting', u'practice',
u'conditions', u'mentoring', u'intrests', u'determination', u'uses', u'praying',
u'confidentiality', u'guaranteed', u'creation', u'lasts', u'authority', u'adults', u'department',
u'structure', u'virtue', u'positions', u'rotten', u'danger', u'remains', u'agreeable',
u'unfaithful', u'sought', u'mentioned', u'measure', u'confess', u'humiliating', u'designed',
u'prospective', u'lists', u'connected', u'consequences', u'earned', u'humans', u'luckiest',
u'approval', u'limits', u'addition', u'diversity', u'swinging', u'capture', u'unconventional',
u'appearances', u'effect', u'reflection', u'distant', u'skill', u'burden', u'limitations', u'primal',
u'secretive', u'cruel', u'tick', u'performed', u'presents', u'connects', u'vast', u'solution',
u'babygirl', u'evolving', u'suitable', u'rights', u'involve', u'gratification', u'indulge',
u'grammar', u'principles', u'crucial', u'fairytale', u'deserving', u'orgasmic', u'suggested',
u'exception', u'tastes', u'iq', u'entitled', u'practices', u'wiser', u'identify', u'candidate',
u>manual', u'subs', u'souls', u'technique', u'item', u'priceless', u'merely', u'wealth',

u'scenarios', u'beholder', u'policy', u'main', u'relive', u'language', u'relative', u'topics',
u'consent', u'awareness', u'secondary', u'worrying', u'stigma', u'according', u'associated',
u'offered', u'different', u'assist', u'solve', u'admitting', u'audience', u'critical', u'expressing',
u'violence', u'thrive', u'general', u'imagination', u'revolve', u'assets', u'forbidden',
u'creatures', u'everlasting', u'abandon', u'experiance', u'acts', u'rapport', u'sacred', u'bloom',
u'suffer', u'idea', u'circumstance', u'embraces', u'compensated', u'mentality', u'part',
u'contrary', u'tale', u'arouse', u'submissives', u'positivity', u'complement', u'reverse',
u'consume', u'creativity', u'itself', u'emphasize', u'craved', u'alluring', u'womanly',
u'follows', u'foolish', u'connecting', u'directed', u'possession', u'constant', u'reputation',
u'utterly', u'failures', u'power', u'safer', u'talent', u'punishments', u'trait', u'praise',
u'attachment', u'sadistic', u'addiction', u'despite', u'defend', u'deprived', u'completes',
u'compensation', u'aforementioned', u'timing', u'lifestyles', u'yearning', u'realization',
u'empathy', u'earning', u'tremendous', u'certain', u'arises', u'curiosity', u'domestic',
u'ideals', u'proper', u'simpler', u'seriousness', u' blessings', u'rushed', u'biggest', u'limited',
u'risk', u'reciprocal', u'emptiness', u'christ', u'conduct', u'humility', u'punishment',
u'abundance', u'expects', u'carnal', u'supporter', u'perception', u'associate', u'eternal',
u'primary', u'especially', u'roles', u'thoroughly', u'thorough', u'genital', u'coin', u'yang',
u'solutions', u'adoration', u'fluid', u'capacity', u'adding', u'brutal', u'forceful', u'arousing',
u'vices', u'humanity', u'companions', u'surround', u'dedicate', u'accomplishments',
u'differences', u'admiration', u'submitting', u'eternity', u'comforting', u'femininity',
u'entertain', u'suggestions', u'unconditional', u'vital', u'unbelievable', u'paramount',
u'insecurities', u'feature', u'lustful', u'equality', u'intercourse', u'factors', u'hypnosis',
u'outcome', u'quantity', u'believing', u'expectation', u'regarding', u'discussing',
u'commitments', u'rekindle', u'endeavor', u'blend', u' blessing', u'guided', u'gentleness',
u'except', u'kills', u'accepted', u'sanity', u'provides', u'of', u'regard', u'strongly', u'jaded',

u'includes', u'vibes', u'debates', u'equation', u'challenging', u'simplicity', u'weaknesses',
u'developed', u'generation', u'experimenting', u'dislikes', u'inhibitions', u'imperfect',
u'nagging', u'risks', u'added', u'grace', u'achieved', u'degrading', u'spells', u'succeed',
u'prescription', u'context', u'overlook', u'political', u'demand', u'behavior', u'disappointing',
u'abilities', u'exclusively', u'seperate', u'admittedly', u'seemingly', u'togetherness',
u'mundane', u'worthwhile', u'bonding', u'karma', u'unavailable', u'expanding', u'sort',
u'struggles', u'complicate', u'profound', u'cerebral', u'chaos', u'regards', u'which', u'purely',
u'highest', u'vanilla', u'encouragement', u'wonders', u'growth', u'nationalities', u'universe',
u'perks', u'goodness']

50: JOBS & EDUCATION

[u'military', u'professor', u'tech', u'debt', u'worker', u'professionally', u'upscale',
u'corporate', u'working', u'assistant', u'artist', u'masseur', u'former', u'engineering',
u'majoring', u'industry', u'graduate', u'full', u'writer', u'currently', u'community', u'ba',
u'athlete', u'licensed', u'degrees', u'owner', u'enforcement', u'studying', u'phd',
u'construction', u'chef', u'established', u'mechanic', u'law', u'certified', u'graduated',
u'profession', u'teacher', u'marine', u'vet', u'trainer', u'veteran', u'justice', u'culinary',
u'gainfully', u'farm', u'healthcare', u'masters', u'attending', u'management', u'employed',
u'grad', u'career', u'owns', u'army', u'medical', u'manager', u'therapist', u'finishing',
u'associates', u'students', u'businesses', u'education', u'biz', u'homeowner', u'volunteer',
u'executive', u'field', u'works', u'firefighter', u'school', u'fulltime', u'advanced', u'officer',
u'amateur', u'coach', u'nursing', u'software', u'engineer', u'college', u'pilot', u'driver',
u'psychology', u'christian', u'disabled', u'health', u'traveled', u'traveler', u'bachelors',

u'retired', u'workers', u'contractor', u'degree', u'musician', u'attend', u'catholic', u'class',
u'student', u'english']

51: SEX TERMS & BODY PARTS

[u'hole', u'bust', u'played', u'smell', u'loads', u'suck', u'bite', u'dicks', u'bubble',
u'tight', u'sucked', u'load', u'pound', u'bj', u'sweaty', u'balls', u'horny', u'phat', u'pussy',
u'nuts', u'licked', u'squirt', u'ass', u'juicy', u'eaten', u'completion', u'wet', u'grind', u'cocks',
u'tits', u'stiff', u'booty', u'luv', u'dry', u'meat', u'milk', u'creamy', u'fuck', u'butt', u'lick',
u'dick', u'cock', u'drain', u'unload', u'drained', u'breast', u'cum', u'blowjob', u'pounded',
u'deepthroat', u'throat', u'swallow', u'fucked', u'asses', u'nipples', u'nut', u'sloppy', u'stroke',
u'boobs', u'hungry', u'blow', u'taste', u'feed', u'rim', u'stick', u'breasts', u'clit', u'asshole',
u'titties', u'bbc']

52: LOCATIONS

[u'sioux', u'rd', u'boston', u'orleans', u'east', u'raised', u'oregon', u'originally', u'sw',
u'sc', u'city', u'san', u'madison', u'carolina', u'west', u'tx', u'tn', u'moines', u'paso', u'wayne',
u'palm', u'santa', u'colorado', u'richmond', u'houston', u'los', u'ohio', u'ca', u'louisiana',
u'atlanta', u'va', u'columbia', u'angeles', u'francisco', u'arkansas', u'indiana', u'las', u'denver',
u'dc', u'cape', u'born', u'columbus', u'vegas', u'live', u'southern', u'pittsburgh', u'america',
u'nearby', u'central', u'area', u'moved', u'salem', u'louis', u'idaho', u'fe', u'fl', u'relocating',
u'grew', u'charlotte', u'philly', u'northern', u'lived', u'albany', u'grand', u'cedar', u'within',

u'kansas', u'ga', u'relocated', u'located', u'fresno', u'university', u'austin', u'montana', u'sf',
u'michigan', u'jersey', u'savannah', u'utah', u'lincoln', u'washington', u'north', u'diego',
u'georgia', u'county', u'cali', u'coast', u'bay', u'springs', u'near', u'mexico', u'seattle', u'fort',
u'tucson', u'downtown', u'nyc', u'texas', u'jackson', u'california', u'macon', u'miles',
u'august', u'island', u'des', u'portland', u'ky', u'visiting', u'nashville', u'baltimore', u'chicago',
u'alaska', u'south', u'dallas', u'areas', u'orange', u'antonio', u'ny', u'dakota', u'memphis',
u'surrounding', u'tampa', u'md', u'ms', u'florida', u'iowa', u'western', u'kentucky',
u'alabama', u'orlando', u'nj', u'nc', u'nw', u'hawaii', u'port', u'local', u'springfield', u'myrtle',
u'oklahoma', u'reno', u'falls', u'state', u'cities', u'virginia', u'valley', u'missouri', u'arizona',
u'wisconsin', u'reside', u'pa', u'moving', u'jacksonville', u'brooklyn', u'phoenix', u'relocate',
u'metro', u'usa', u'york', u'europe', u'miami', u'tennessee', u'illinois']

53: RESPONSE

[u'message', u'listed', u'likely', u'resonates', u'letter', u'saying', u'clicked',
u'respond', u'asked', u'bothers', u'wish', u'sure', u'exact', u'leave', u'flagging', u'typing',
u'note', u'anyways', u'dare', u'pass', u'section', u'replied', u'sparked', u'read', u'this', u'fits',
u'robot', u'bother', u'appealing', u'clearly', u'hope', u'handle', u'youll', u'pleasantly',
u'convince', u'chance', u'cant', u'disappoint', u'responded', u'requests', u'speak', u'asking',
u'post', u'responding', u'interested', u'flag', u'regret', u'prove', u'write', u'remember',
u'intrigues', u'drop', u'intrigued', u'directly', u'rambling', u'inquire', u'describe', u'you',
u'answers', u'excuse', u'bless', u'ask', u'kindly', u'catches', u'describes', u'look', u'reads',
u'describing', u'telling', u'hesitate', u'anyway', u'written', u'piqued', u'page', u'tell',
u'posting', u'answer', u'explain', u'stop', u'flagged', u'interest', u'if', u'youre', u'delete',

u'assume', u'answered', u'wait', u'thank', u'interesting', u'advance', u'question', u'sounds',
u'answering', u'shot', u'considering', u'sound', u'peaked', u'carefully', u'understand', u'bill',
u'disappointed', u'forward', u'profile', u'mention', u'legit', u'assure', u'ad', u'add', u'match',
u'assuming', u'replying', u'guarantee', u'peaks', u'appeals', u'hearing', u'quickly', u'criteria',
u'clicking', u'dissappointed', u'indeed', u'considered', u'follow', u'glad', u'instructions', u'feel',
u'wrote', u'change', u'questions', u'ignore', u'chose', u'offends', u'from', u'choose',
u'chances', u'responds', u'spammer', u'contacting', u'promise']

A.3. PERCEPTION SURVEY QUESTIONNAIRE

Q19 Survey on text and author perception: cover letter

This survey is part of a study of linguistic variation in English. It should take less than 3 minutes to complete.

Your participation will help us better understand the meaning of various language features.

No data besides answers to the survey questions will be collected.

Mechanical Turk worker IDs will only be collected for the purposes of distributing compensation and will not be associated with survey responses.

The Principal Investigator for this study is Prof. Lars Hinrichs, Dept. of English, University of Texas at Austin, (512) 471-8755.

This study has been approved by The University of Texas' Institutional Review Board, study number 2016-01-0055, and Documentation of Consent has been waived.

You may contact the IRB at (512) 471-8871.

Thank you for your participation!

Q21 The following text is intended to be posted on a social networking web site. Please read it and answer the questions on the next page. (Identifying information such as email addresses has been removed.)Thank you!

Q31 Hi there, I'm looking for fun people to hang out with from time to time, especially on the weekends. Going to the beach, going to the movies or having a drink. Or just explore the city. If ur interested, contact me at _____@gmail.com for more information. Looking forward to hearing from u!

Q32 The author is ...

- male
- female

Q2 The author is ...

- homosexual
- heterosexual

Q23 The author is ...

- Very friendly
- Friendly
- Somewhat friendly
- Somewhat unfriendly
- Very unfriendly

Q4 The author is writing for a ...

- man
- woman

Q8 The author seems ...

- Very sensitive
- Sensitive
- Somewhat sensitive
- Insensitive
- Very insensitive

Q10 I'd guess the author is ...

- Black / African American
- Asian
- White
- Hispanic / Latino

Q12 The author seems ...

- Very assertive
- Assertive
- Somewhat assertive
- Somewhat timid
- Timid

Q25 The author seems ...

- Very attractive
- Attractive
- Somewhat attractive
- Unattractive
- Very unattractive

Q14 The author seems ...

- Very educated
- Somewhat educated
- Of average education
- Somewhat uneducated
- Very uneducated

Q26 Men are ...

Q27 Women are ...

Q24 Would you reply to this ad?

- Yes, very likely
- Likely
- Somewhat likely
- Unlikely
- No, very unlikely

Q23 Do you have any further comments? Please share!

Q16 Thanks! To finish up, please answer a few questions about yourself.

Q18 I am ...

- male
- female
- other, please specify _____

Q20 I consider myself ...

- White
- Hispanic / Latino
- Black / African-American
- Asian
- other, please specify _____

Q22 My age:

Q24 I consider myself ...

- heterosexual
- homosexual
- bisexual
- other, please specify _____

Q26 I grew up in this city and this state:

Q28 I now live in this city:

Q30 Thanks for your input! Make sure to copy the code provided here to collect your salary on Mechanical Turk. You will input it through Mechanical Turk to indicate your completion of the study. Then click the button on the bottom of the page to submit your answers. You will not receive credit unless you click this button. `{e://Field/mTurkCode}`

Works Cited

- Acker, Joan. 1973. Women and social stratification: A case of intellectual sexism. *American Journal of Sociology* 78(4). 936–945.
- Aggarwal, Charu C., Alexander Hinneburg & Daniel A. Keim. 2001. On the surprising behavior of distance metrics in high dimensional space. *Database Theory: ICDT 2001*, 420–434. (Lecture Notes in Computer Science 1973). Berlin: Springer.
- Agha, Asif. 2003. The social life of cultural value. *Language & Communication* 23(3–4). 231–273.
- Agha, Asif. 2007. *Language and social relations*. Cambridge: Cambridge University Press.
- Androutsopoulos, Jannis. 2014. Linguaging when contexts collapse: Audience design in social networking. *Discourse, Context & Media* 4–5. 62–73.
- Aronovitch, Charles D. 1976. The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. *The Journal of Social Psychology* 99(2). 207–220.
- Austin, J. L. 1962. *How to do things with words*. (William James Lectures 1955). Cambridge: Harvard University Press.
- Baayen, Harald, Richard Piepenbrock & Hedderick van Rijn. 1993. The Celex database on CD-ROM. *Linguistic Data Consortium. Philadelphia, PA*.
- Baker, Paul. 2014. *Using corpora to analyze gender*. London: Bloomsbury Academic.
- Bakir, Murtadha. 1986. Sex differences in the approximation to standard Arabic: a case study. *Anthropological Linguistics* 28(1). 3–9.
- Ball, Peter. 1983. Stereotypes of Anglo-Saxon and non-Anglo-Saxon accents: Some exploratory Australian studies with the matched guise technique. *Language Sciences* 5(2). 163–183.
- Bamman, David, Jacob Eisenstein & Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2). 135–160.
- Baron, Naomi S. 2002. *Alphabet to email: How written English evolved and where it's heading*. New York: Routledge.
- Bell, Allan. 1982. Radio: The style of news language. *Journal of Communication* 32(1). 150–164.
- Bell, Allan. 1984. Language style as audience design. *Language in Society* 13(2). 145–204.
- Bell, Allan. 1992. Hit and miss: Referee design in the dialects of New Zealand television advertisements. *Language & Communication* 12(3). 327–340.
- Bell, Allan. 2001. Back in style: Reworking audience design. In Penelope Eckert & John R. Rickford (eds.), *Style and sociolinguistic variation*, 139–169. Cambridge: Cambridge University Press.
- Bensinger, Greg. 2017. Investors try to tap into the next Craigslist, regardless of earnings. *Wall Street Journal*, sec. Tech. <http://www.wsj.com/articles/investors-try-to-tap-into-the-next-craigslist-regardless-of-earnings-1484654407> (18 January, 2017).

- Bergen, David J. & John E. Williams. 1991. Sex stereotypes in the United States revisited: 1972–1988. *Sex Roles* 24(7–8). 413–423.
- Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
- Bilous, Frances R. & Robert M. Krauss. 1988. Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads. *Language & Communication* 8(3). 183–194.
- Blatt, Ben. 2017. *Nabokov's favorite word is mauve: What the numbers reveal about the classics, bestsellers, and our own writing*. Kindle Edition. New York: Simon & Schuster.
- Boland, Julie E. & Robin Queen. 2016. If your house is still available, send me an email: Personality influences reactions to written errors in email messages. *Plos One* 11(3). 1–17.
- Bourdieu, Pierre & Luc Boltanski. 1975. Le fétichisme de la langue. *Actes de la recherche en sciences sociales* 1(4). 2–32.
- Bradley, John. 1988. Yanyuwa: “Men speak one way, women speak another.” *Aboriginal Linguistics* 1. 126–34.
- Britain, David. 1992. Linguistic change in intonation: The use of high rising terminals in New Zealand English. *Language Variation and Change* 4(1). 77–104.
- Broverman, Inge K., Susan Raymond Vogel, Donald M. Broverman, Frank E. Clarkson & Paul S. Rosenkrantz. 1972. Sex-role stereotypes: A current appraisal. *Journal of Social Issues* 28(2). 59–78.
- Brown, Penelope & Stephen C. Levinson. 1978. Universals in language usage: Politeness phenomena. In Ester N. Goody (ed.), *Questions and politeness: Strategies in social interaction*, 56–311. Cambridge: Cambridge University Press.
- Brownlow, Sheila, Julie A. Rosamond & Jennifer A. Parker. 2003. Gender-linked linguistic behavior in television interviews. *Sex Roles* 49(3–4). 121–132.
- Bucholtz, Mary. 1996. Geek the girl: Language, femininity, and female nerds. *Gender and belief systems: proceedings of the fourth Berkeley women and language conference*, 119–132. Berkeley: Berkeley women and language group.
- Bucholtz, Mary. 1999a. You da man: Narrating the racial other in the production of white masculinity. *Journal of Sociolinguistics* 3(4). 443–460.
- Bucholtz, Mary. 1999b. “Why be normal?”: Language and identity practices in a community of nerd girls. *Language in Society* 28(2). 203–223.
- Bucholtz, Mary & Kira Hall. 2004. Theorizing identity in language and sexuality research. *Language in Society* 33(4). 469–515.
- Bumble, Inc. 2017. Bumble - Help. *Bumble*. <https://bumble.com/en-us/faq> (20 March, 2017).
- Burger, John D., John Henderson, George Kim & Guido Zarrella. 2011. Discriminating gender on Twitter. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1301–1309. Association for Computational Linguistics.
- Butler, Judith. 1999. *Gender trouble: Feminism and the subversion of identity*. 2nd ed. New York: Routledge.

- Cameron, Deborah. 1992a. "Not gender difference but the difference gender makes" - explanation in research on sex and language. *International Journal of the Sociology of Language* 94. 13–26.
- Cameron, Deborah. 1992b. Review of Tannen (1990). *Feminism & Psychology* 2(3). 465–468.
- Cameron, Deborah. 1997. Performing gender identity: Young men's talk and the construction of heterosexual masculinity. In Sally Johnson & Ulrike Hanna Meinhof (eds.), *Language and Masculinity*, 47–64. Oxford: Blackwell.
- Cameron, Deborah (ed.). 1998a. *The feminist critique of language: A reader*. 2nd ed. New York: Routledge.
- Cameron, Deborah. 1998b. Gender, language, and discourse: A review essay. *Signs* 23(4). 945–973.
- Cameron, Deborah. 2009. Sex/gender, language and the new biologism. *Applied Linguistics* 31(2). 173–92.
- Cameron, Deborah & Jennifer Coates. 1985. Some problems in the sociolinguistic explanation of sex differences. *Language & Communication* 5(3). 143–151.
- Cameron, Deborah, Fiona McAlinden & Kathy O'Leary. 1988. Lakoff in context: The social and linguistic functions of tag questions. In Jennifer Coates & Deborah Cameron (eds.), *Women in their speech communities*, 74–93. New York: Longman.
- Campbell-Kibler, Kathryn. 2008. I'll be the judge of that: Diversity in social perceptions of (ING). *Language in Society* 37(5). 637–659.
- Campbell-Kibler, Kathryn. 2009. The nature of sociolinguistic perception. *Language Variation and Change* 21(1). 135–156.
- Campbell-Kibler, Kathryn. 2010. Sociolinguistics and perception. *Language and Linguistics Compass* 4(6). 377–389.
- Campbell-Kibler, Kathryn. 2011. Intersecting variables and perceived sexual orientation in men. *American Speech* 86(1). 52–68.
- Cannon, Garland. 1989. Abbreviations and acronyms in English word-formation. *American Speech* 64(2). 99–127.
- Chafe, Wallace. 1988. Punctuation and the prosody of written language. *Written Communication* 5(4). 395–426.
- Chambers, Jack K. 2009. *Sociolinguistic theory*. 2nd, revised ed. (Language in Society 22). Chichester: Wiley-Blackwell.
- Chambers, Jack K. & Peter Trudgill. 1998. *Dialectology*. 2nd ed. (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.
- Champely, Stephane, Claus Ekstrom, Peter Dalgaard, Jeffrey Gill, Stephan Weibelzahl, Aditya Anandkumar, Clay Ford, Robert Volcic & Helios De Rosario. 2017. *pwr: Basic functions for power analysis*. <https://cran.r-project.org/web/packages/pwr/index.html> (7 April, 2017).
- Cheshire, Jenny. 1982. *Variation in an English dialect: A sociolinguistic study*. . Vol. 37. (Cambridge Studies in Linguistics). Cambridge: Cambridge University Press.
- Cheshire, Jenny. 2002. Sex and gender in variationist research. In Jack K. Chambers & Natalie Schilling (eds.), *The handbook of language variation and change*, 423–43. Malden, MA: Wiley and Sons.

- Clarke, Sandra, Ford Elms & Amani Youssef. 1995. The third dialect of English: Some Canadian evidence. *Language Variation and Change* 7(2). 209–228.
- Coates, Jennifer. 1989. Gossip revisited: language in all-female groups. In Jennifer Coates & Deborah Cameron (eds.), *Women in their speech communities*, 94–122. New York: Longman.
- Coates, Jennifer. 1991. *Women talk: Conversation between women friends*. Wiley-Blackwell.
- Coates, Jennifer. 2004. *Women, men and language*. 3rd ed. (Studies in Language and Linguistics). Harlow: Pearson Education.
- Cohen, Jacob. 1992. Statistical power analysis. *Current directions in psychological science* 1(3). 98–101.
- Coulmas, Florian. 2013. *Sociolinguistics: The study of speakers' choices*. Cambridge: Cambridge University Press.
- Coupland, J. 2000. Past the “perfect kind of age”? Styling selves and relationships in over-50s dating advertisements. *Journal of Communication* 50(3). 9–30.
- Coupland, Justine. 1996. Dating advertisements: Discourses of the commodified self. *Discourse & Society* 7(2). 187–207.
- Craigslist.org. 2016. craigslist | about > factsheet. <https://www.craigslist.org/about/factsheet> (19 February, 2016).
- Crawley, Michael J. 2007. *The R book*. 1st ed. Chichester: John Wiley & Sons.
- Crystal, David. 2011. *Dictionary of linguistics and phonetics*. Chichester: John Wiley & Sons.
- Danescu-Niculescu-Mizil, Cristian, Robert West, Dan Jurafsky, Jure Leskovec & Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. *Proceedings of the 22nd international conference on World Wide Web*, 307–318. ACM. <http://dl.acm.org/citation.cfm?id=2488416> (15 April, 2017).
- Danet, Brenda. 1998. Text as mask: Gender, play, and performance on the Internet. In SG Jones (ed.), *CyberSociety 2.0: Revisiting computer-mediated communication and community*. Thousand Oaks: Sage Publications.
- Davis, Simon. 1990. Men as success objects and women as sex objects: A study of personal advertisements. *Sex Roles* 23(1–2). 43–50.
- Delphy, Christine. 1981. Women in stratification studies. In Helen Roberts (ed.), *Doing feminist research*, 114–128. London: Routledge & Kegan Paul.
- Denning, Keith M. 1987. *Variation in language: NWAV-XV at Stanford*. Department of Linguistics, Stanford University.
- Denno, Deborah. 1982. Sex differences in cognition: A review and critique of the longitudinal evidence. *Adolescence*(17). 779–88.
- Deuchar, Margaret. 1988. A pragmatic account of women’s use of standard speech. In Jennifer Coates & Deborah Cameron (eds.), *Women in their speech communities*, 27–32. London: Longman.
- Dixon, Roland B. & Alfred L. Kroeber. 1903. The native languages of California. *American Anthropologist* 5(1). 1–26.
- Drager, Katie. 2013. Experimental methods in sociolinguistics. In Janet Holmes & Kirk Hazen (eds.), *Research methods in sociolinguistics: A practical guide*, 58–73. Hoboken: Wiley-Blackwell.

- Duda, Richard O., Peter E. Hart & David G. Stork. 2012. *Pattern classification*. 2nd ed. Chichester: John Wiley & Sons.
- Eaton, Asia Anna & Suzanna Rose. 2011. Has dating become more egalitarian? A 35 year review using sex roles. *Sex Roles* 64(11–12). 843–862.
- Eckert, Penelope. 1989. *Jocks and burnouts: social categories and identity in the high school*. New York: Teachers College Press.
- Eckert, Penelope. 1990. The whole woman: sex and gender differences in variation. *Language Variation and Change* 1(3). 245–267.
- Eckert, Penelope. 1996. Vowels and nail polish: the emergence of linguistic style in the preadolescent heterosexual marketplace. In Natasha Werner, Jocelyn Ahlers, Leela Bilmes, Monica Oliver, Suzanne Wertheim & Melinda Chen (eds.), *Gender and belief systems: proceedings of the fourth Berkeley women and language conference*, vol. 183–90, 190. Berkeley: Berkeley Women and Language Group.
- Eckert, Penelope. 2000. *Linguistic variation as social practice: the linguistic construction of identity in Belten High*. (Language in Society 27). Malden: Blackwell.
- Eckert, Penelope. 2008a. Variation and the indexical field. *Journal of Sociolinguistics* 12(4). 453–476.
- Eckert, Penelope. 2008b. Variation and the indexical field. *Journal of sociolinguistics* 12(4). 453–476.
- Eckert, Penelope. 2012a. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology* 41. 87–100.
- Eckert, Penelope. 2012b. Coding for gender and sexuality. Portland.
- Eckert, Penelope. 2014. The problem with binaries: Coding for gender and sexuality. *Language and Linguistics Compass* 8(11). 529–535.
- Eckert, Penelope & Sally McConnell-Ginet. 1992. Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology* 21. 461–490.
- Eckert, Penelope & Sally McConnell-Ginet. 2013. *Language and gender*. 2nd ed. Cambridge: Cambridge University Press.
- Eckert, Penelope & John R. Rickford. 2001. *Style and sociolinguistic variation*. Cambridge: Cambridge University Press.
- Edelsky, Carole. 1979. Question intonation and sex roles. *Language in Society* 8(1). 15–32.
- Edwards, John. 2013. *Sociolinguistics: A very short introduction*. Oxford: Oxford University Press.
- Ehrlich, Susan & Miriam Meyerhoff. 2014. Introduction: language, gender, and sexuality. In Susan Ehrlich, Miriam Meyerhoff & Janet Holmes (eds.), *The handbook of language, gender, and sexuality*, 1–20. 2nd ed. Malden: Wiley-Blackwell.
- Ehrlich, Susan, Miriam Meyerhoff & Janet Holmes (eds.). 2014. *The handbook of language, gender, and sexuality*. 2nd ed. Chichester: Wiley-Blackwell.
- Eisenstein, Jacob. 2017. Identifying regional dialects in online social media. In Charles Boberg, John Nerbonne & Dominic Watt (eds.), *Handbook of dialectology*. New York: Wiley.

- Ellison, Nicole, Rebecca Heino & Jennifer Gibbs. 2006. Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication* 11(2). 415–441.
- Feagin, Crawford. 1979. *Variation and change in Alabama English: A sociolinguistic study of the white community*. Washington D.C.: Georgetown University Press.
- Fein, Ellen & Sherrie Schneider. 1995. *The rules: Time-tested secrets for capturing the heart of Mr. Right*. London: HarperCollins.
- Fischer, John L. 1958. Social influences on the choice of a linguistic variant. *Word* 14(1). 47–56.
- Fishman, Pamela M. 1983. Interaction: The work women do. In Barrie Thorne, Cheris Kramarae & Nancy Henley (eds.), *Language, gender, and society*, 89–103. Rowley: Newbury House.
- Freed, Alice. 1992. We understand perfectly: A critique of Tannen's view of cross-sex communication. In Kira Hall, Mary Bucholtz & B Moonwomon (eds.), *Locating power: Proceedings of the second Berkeley women and language conference*, 144–152. Berkeley: Berkeley Women and Language Group.
- Fullick, Melonie. 2013. "Gendering" the self in online dating discourse. *Canadian Journal of Communication* 38(4). 545–562.
- Gal, Susan. 1978. Peasant men can't get wives: Language change and sex roles in a bilingual community. *Language in Society* 7(1). 1–16.
- Gauchat, L. 1905. *L'unité phonétique dans le patois d'une commune*. Halle: Max Niemeyer.
- Giles, Howard & Andrew C. Billings. 2004. Assessing language attitudes: Speaker evaluation studies. In Allen Davies & Catherine Elder (eds.), *The handbook of applied linguistics*, 187–210. Oxford: Blackwell.
- Giles, Howard, Nikolas Coupland & Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*, 1–69. Cambridge: Cambridge University Press.
- Gilliéron, Jules & Edmond Edmont. 1902. *Atlas linguistique de la France*. Paris: Champion.
- Goodwin, Marjorie Harness. 1980. Directive-response speech sequences in girls' and boys' task activities. In Sally McConnell-Ginet, Ruth Borker & Nelly Furman (eds.), *Women and language in literature and society*, 157–74. New York: Praeger.
- Gordon, Matthew & Jeffrey Heath. 1998. Sex, sound symbolism, and sociolinguistics. *Current Anthropology* 39(4). 421–449.
- Greenwood, Shannon, Andrew Perrin & Maeve Duggan. 2016. *Social Media Update 2016*. Washington, D.C.: Pew Research Center.
<http://www.pewinternet.org/2016/11/11/social-media-update-2016/> (11 April, 2017).
- Grimmer, Justin & Gary King. 2011. A general purpose computer-assisted clustering methodology. *Proceedings of the National Academy of Sciences* 108(7). 2643–2650.
- Guiller, Jane & Alan Durndell. 2007. Students' linguistic behaviour in online discussion groups: Does gender matter? *Computers in Human Behavior* 23(5). 2240–2255.
- Haas, Mary R. 1944. Men's and women's speech in Koasati. *Language* 20(3). 142–149.

- Hall, Kira. 1995. Lip service on the fantasy lines. In Kira Hall & Mary Bucholtz (eds.), *Gender articulated: Language and the socially constructed self*, 183–216. New York: Routledge.
- Hall, Kira. 1999. Performativity. *Journal of Linguistic Anthropology* 9(1–2). 184–187.
- Hall, Kira & Veronica O'Donovan. 1996. Shifting gender positions among Hindi-speaking hijras. In Victoria Bergvall, Janet M Bing & Alice F Freed (eds.), *Rethinking language and gender research: Theory and practice*, 228–66. London: Longman.
- Halpern, Diane F. 1986. *Sex differences in cognitive abilities*. Hillsdale: Erlbaum.
- Harrington, Kate, Lia Litosseliti, Helen Sauntson & Jane Sunderland (eds.). 2008. *Gender and language research methodologies*. Houndmills, Basingstoke, Hampshire ; New York: Palgrave Macmillan.
- Heilbrun, Carolyn G. & Katha Politt. 2008. *Writing a woman's life*. New York: Norton.
- Henry, Victor. 1879. Sur le parler des hommes et le parler des femmes dans la langue chiquita. *Revue de Linguistique et de Philologie Comparee* 12. 305–313.
- Herring, Susan. 1999. Interactional coherence in CMC. *Journal of Computer-Mediated Communication* 4(4). (4 April, 2017).
- Herring, Susan C. 2012. Grammar and electronic communication. *The encyclopedia of applied linguistics*, 1–7. Oxford: Blackwell.
- Herring, Susan C & John C Paolillo. 2006. Gender and Genre Variation in Weblogs. *Journal of Sociolinguistics* 10(4). 439–459.
- Herring, Susan C & John C. Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics* 10(4). 439–459.
- Heylighen, Francis & Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science* 7(3). 293–340.
- Hitsch, Günter J., Ali Hortaçsu & Dan Ariely. 2010. What makes you click? Mate preferences in online dating. *Quantitative Marketing and Economics* 8(4). 393–427.
- Hogg, Michael A. 1985. Masculine and feminine speech in dyads and groups: A study of speech style and gender salience. *Journal of Language and Social Psychology* 4(2). 99–112.
- Holmes, Janet. 1990. Hedges and boosters in women's and men's speech. *Language & Communication* 10(3). 185–205.
- Holmes, Janet. 1993. New Zealand women are good to talk to: An analysis of politeness strategies in interaction. *Journal of Pragmatics* 20(2). 91–116.
- Holmes, Janet & David Britain. 1998. Comment on sex, sound symbolism and sociolinguistics. *Current Anthropology* 39(4). 442.
- Holmes, Janet & Miriam Meyerhoff (eds.). 2003. *The handbook of language and gender*. 1st ed. Malden: Blackwell.
- Holmquist, Jonathan C. 1985. Social correlates of a linguistic variable: A study in a Spanish village. *Language in Society* 14(2). 191–203.
- Horvath, Barbara M. 1985. *Variation in Australian English: The sociolects of Sydney*. (Cambridge Studies in Linguistics 45). Cambridge: Cambridge University Press.
- Hudson, Richard Anthony. 1996. *Sociolinguistics*. 2nd ed. (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

- Huygens, Ingrid & Graham M. Vaughan. 1983. Language attitudes, ethnicity and social class in New Zealand. *Journal of Multilingual & Multicultural Development* 4(2–3). 207–223.
- Hyde, Janet & Marcia Linn. 1988. Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin* 104(1). 53–69.
- Ibrahim, Muhammad H. 1986. Standard and prestige language: A problem in Arabic sociolinguistics. *Anthropological Linguistics* 28(1). 115–126.
- Jagger, Elizabeth. 1998. Marketing the self, buying an other: Dating in a post modern, consumer society. *Sociology* 32(4). 795–814.
- James, Deborah. 1996. Women, men and prestige speech forms: A critical review. In Victoria Bergvall, Janet M Bing & Alice F Freed (eds.), *Rethinking language and gender research: Theory and practice*, 98–125. New York: Addison Wesley Longman.
- Janssen, Anna & Tamar Murachver. 2004. The Relationship between Gender and Topic in Gender-Preferential Language Use. *Written Communication* 21(4). 344–367.
- Jawad, Abdel. 1987. Cross-dialectal variation in Arabic: competing prestige forms. *Language in Society* 16(3). 359–67.
- Jespersen, Otto. 1964. *Language: its nature and development and origin*. London: George Allen & Unwin.
- Jones, Elizabeth S., Cynthia Gallois, Victor J. Callan & Michelle Barker. 1995. Language and power in an academic context: The effects of status, ethnicity, and sex. *Journal of Language and Social Psychology* 14(4). 434–461.
- Kelly, Ryan. 2016. *PyEnchant. Spellchecking library for Python*. <http://www.rfk.id.au/software/pyenchant/>.
- Kemp, William & Malcah Yaeger-Dror. 1991. Changing realizations of a in (a)tion in relation to the front a-back a opposition in Quebec French. In Penelope Eckert (ed.), *New ways of analyzing sound change*, 127–84. San Diego: Academic Press.
- Khan, Farhat. 1991. Final consonant cluster simplification in a variety of Indian English. In Jenny Cheshire (ed.), *English around the world: Sociolinguistic perspectives*, 288–298. Cambridge: Cambridge University Press.
- Kiesling, Scott Fabius. 1998. Men's identities and sociolinguistic variation: The case of fraternity men. *Journal of Sociolinguistics* 2(1). 69–99.
- Kiesling, Scott Fabius. 2007. Men, masculinities, and language. *Language and Linguistics Compass* 1(6). 653–673.
- Kiesling, Scott Fabius. 2009. Fraternity men: variation and discourses of masculinity. In Nikolas Coupland & Adam Jaworski (eds.), *The new sociolinguistics reader*, 187–200. New York: Palgrave Macmillan.
- Koestner, Richard & Ladd Wheeler. 1988. Self-presentation in personal advertisements: The influence of implicit notions of attraction and role expectations. *Journal of Social and Personal Relationships* 5(2). 149–160.
- Koppel, M., S. Argamon & A.R. Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17(4). 401–412.
- Labov, William. 1966. *The social stratification of English in New York City*. Washington: Center for Applied Linguistics.

- Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
- Labov, William. 1978. Where does the linguistic variable stop? A response to Beatriz Lavandera. *Working Papers in Sociolinguistics* 44. 1–23.
- Labov, William. 1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2(2). 205–254.
- Labov, William. 2001. *Principles of linguistic change: Social factors*. Oxford: Blackwell.
- Labov, William. 2006. *The social stratification of English in New York city*. 2nd ed. Cambridge: Cambridge University Press.
- Labov, William, Sharon Ash & Charles Boberg. 2006. *Atlas of North American English: Phonology and Phonetics*. Berlin: Mouton de Gruyter.
- Lakoff, Robin. 1973. Language and woman's place. *Language in Society* 2(1). 45–79.
- Lakoff, Robin. 2004. *Language and woman's place: text and commentaries*. (Ed.) Mary Bucholtz. (Studies in Language and Gender). Oxford: Oxford University Press.
- Lambert, Wallace E., Richard C. Hodgson, Robert C. Gardner & Samuel Fillenbaum. 1960. Evaluational reactions to spoken languages. *The Journal of Abnormal and Social Psychology* 60(1). 44–51.
- Lance, Larry M. 1998. Gender differences in heterosexual dating: A content analysis of personal ads. *The Journal of Men's Studies* 6(3). 297–305.
- Laner, Mary Riege & Nicole A Ventrone. 2000. Dating scripts revisited. *Journal of Family Issues* 21(4). 488–500.
- Larson, Karen A. 1982. Role playing and the real thing: Socialization and standard speech in Norway. *Journal of Anthropological Research* 38(4). 401–410.
- Levon, Erez. 2007. Sexuality in context: Variation and the sociolinguistic perception of identity. *Language in Society* 36(4). 533–554.
- Lillis, Theresa. 2013. *The sociolinguistics of writing*. Edinburgh: Edinburgh University Press.
- Litosseliti, Lia. 2006. *Gender and language theory and practice*. London: Hodder Arnold.
- Long, Bridget L. 2010. Scripts for online dating: A model and theory of online romantic relationship initiation. Bowling Green State University.
https://etd.ohiolink.edu/!etd.send_file?accession=bgsu1268852623&disposition=attachment (23 March, 2017).
- Lueptow, Lloyd B., Lori Garovich-Szabo & Margaret B. Lueptow. 2001. Social change and the persistence of sex typing: 1974–1997. *Social Forces* 80(1). 1–36.
- Macaulay, Ronald KS. 1978a. The myth of female superiority in language. *Journal of Child Language* 5(2). 353–363.
- Macaulay, Ronald KS. 1978b. Variation and consistency in Glaswegian English. In Peter Trudgill (ed.), *Sociolinguistic Patterns in British English*, 132–143. London: E. Arnold.
- Maccoby, Eleanor E. & Carol Nagy Jacklin. 1974. *The psychology of sex differences*. Stanford: Stanford University Press.
- Maltz, Daniel N. & Ruth A. Borker. 1982. A cultural approach to male-female miscommunication. In J Gumperz (ed.), *Language and social identity*, 159–216. Cambridge: Cambridge University Press.

- Marley, Carol. 2007. Metaphors of identity in dating ads and newspaper articles. *Text & Talk* 27(1). 55–78.
- Marley, Carol. 2008a. Truth values and truth-commitment in interdiscursive dating ads. *Language and Literature* 17(2). 137–154.
- Marley, Carol. 2008b. Assuming identities: The workings of intertextual metaphors in a corpus of dating ads. *Journal of Pragmatics* 40(3). 559–576.
- Maurer, Todd J. & Heather R. Pierce. 1998. A comparison of Likert scale and traditional measures of self-efficacy. *Journal of Applied Psychology* 83(2). 324–29.
- McKay, Susan. 2011. Language and the media. In Rajend Meshtrie (ed.), *The Cambridge Handbook of Sociolinguistics*, 396–413. Cambridge: Cambridge University Press.
- Meshtrie, Rajend, Alida Chevalier & Timothy Dunne. 2015. A regional and social dialectology of the BATH vowel in South African English. *Language Variation and Change* 27(1). 1–30.
- Meyerhoff, Miriam. 2011. *Introducing sociolinguistics*. 2nd ed. Hoboken: Taylor & Francis.
- Meyer-Lübke, Wilhelm, Matteo Bartoli & Giacomo Braun. 1927. *Grammatica storica della lingua italiana e dei dialetti toscani*. 2. ed. Torino: Loescher.
- Mikolov, T. & J. Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*. <http://arxiv.org/pdf/1310.4546v1.pdf>.
- Mills, Sara. 2012. *Gender matters: Feminist linguistic analysis*. London: Equinox.
- Mills, Sara & Louise Mullany. 2011. *Language, gender and feminism: theory, methodology and practice*. New York: Routledge.
- Milroy, James. 1982. Probing under the tip of the iceberg: phonological “normalization” and the shape of the speech community. In Suzanne Romaine (ed.), *Sociolinguistic variation in speech communities*, 26–48. London: Edward Arnold.
- Milroy, James. 1989. The concept of prestige in sociolinguistic argumentation. *York Papers in Linguistics* 13. 215–26.
- Milroy, James. 1992. Social network and prestige arguments in sociolinguistics. In Kingsley Bolton & Helen Kwok (eds.), *Sociolinguistics today: International perspectives*, 146–62. London: Routledge.
- Milroy, James & Lesley Milroy. 1978. Belfast: Change and variation in an urban vernacular. In Peter Trudgill (ed.), *Sociolinguistic patterns in British English*, 19–37. London: E. Arnold.
- Milroy, James & Lesley Milroy. 1993. Mechanisms of change in urban dialects: the role of class, social network and gender. *International Journal of Applied Linguistics* 3(1). 57–77.
- Milroy, Lesley. 1987. *Language and social networks*. 2nd ed. New York: Basil Blackwell.
- Milroy, Lesley. 1988. Review of Variation in Australian English. *Language in Society* 17(4). 577–581.
- Milroy, Lesley. 1992. New perspectives in the analysis of sex differentiation in language. In Kingsley Bolton & Helen Kwok (eds.), *Sociolinguistics today: International perspectives*, 163–79. London: Routledge.
- Milroy, Lesley. 2002. *Authority in language: Investigating standard English*. New York: Routledge.

- Mulac, A. & T.L. Lundell. 1994. Effects of gender-linked language differences in adults' written discourse: Multivariate tests of language effects. *Language and Communication* 14. 299–299.
- Namy, Laura L., Lynne C. Nygaard & Denise Sauerteig. 2002. Gender Differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology* 21(4). 422–432.
- NCES. 2017. Standard 1-5 - NCES Statistical Standards. *Statistical standards: Defining race and ethnicity Data*. https://nces.ed.gov/statprog/2002/std1_5.asp (24 January, 2017).
- Nevalainen, Terttu & Helena Raumolin-Brunberg. 2016. *Historical sociolinguistics: Language change in Tudor and Stuart England*. New York: Routledge.
- Newman, M.L., C.J. Groom, L.D. Handelman & J.W. Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45(3). 211–236.
- Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé & Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics*. http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00258 (27 March, 2017).
- Nguyen, Dong, Dolf Trieschnigg & Theo Meder. 2014. Tweetgenie: Development, evaluation, and lessons learned. Association for Computational Linguistics. <http://doc.utwente.nl/94056/> (15 April, 2017).
- Nichols, Patricia. 1983. Linguistic options and choices for black women in the rural south. In Barrie Thorne, Cheris Kramarae & Nancy Henley (eds.), *Language, gender and society*, 54–68. Rowley, MA: Newbury House.
- Nielsen, Joyce McCarl, Glenda Walden & Charlotte A. Kunkel. 2000. Gendered heteronormativity: Empirical illustrations in everyday life. *The Sociological Quarterly* 41(2). 283–296.
- NLTK Project. 2013. nltk.tag.perceptron — NLTK 3.0 documentation. http://www.nltk.org/_modules/nltk/tag/perceptron.html (18 January, 2017).
- NLTK Project. 2015. nltk.tokenize package — NLTK 3.0 documentation. <http://www.nltk.org/api/nltk.tokenize.html> (20 February, 2016).
- Nowson, Scott, Jon Oberlander & Alastair J. Gill. 2005. Weblogs, genres and individual differences. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, 1666–71.
- O'Barr, William M. & Bowman K. Atkins. 1980. “Womens' language” or “powerless language”? In Sally McConnell-Ginet, Ruth Borker & Nelly Furman (eds.), *Women and language in literature and society*, 93–111. New York: Praeger.
- Ochs, Elinor. 1990. Indexicality and socialization. In James W. Stigler, Richard Shweder & Gilbert Herdt (eds.), *Cultural psychology: Essays on comparative human development*, 287–308. Cambridge: Cambridge University Press.
- Ochs, Elinor. 1992. Indexing gender. In Alessandro Duranti & Charles Goodwin (eds.), *Rethinking context: Language as an interactive phenomenon*, 335–58. (Studies in the Social and Cultural Foundations of Language 11). Cambridge: Cambridge University Press.

- Ohala, John. 1994. The frequency code underlies the sound-symbolic use of voice pitch. In L Hinton, J Nichols & John Ohala (eds.), *Sound symbolism*, 325–47. Cambridge: Cambridge University Press.
- Orton, Harold. 1962. *Survey of English dialects: Introduction*. . Vol. 1. Leeds: EJ Arnold.
- Orton, Harold, Stewart Sanderson & John Widdowson. 1962. *The linguistic atlas of England*. London: Humanities Press.
- Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.
- Peters, Cara O., Jane B. Thomas & Richard Morris. 2013. Looking for love on Craigslist: An examination of gender differences in self-marketing online. *Journal of Marketing Development & Competitiveness* 7(3). 79–95.
- Pew Research Center. 2017. Internet/Broadband Fact Sheet. *Pew Research Center: Internet, Science & Tech*. <http://www.pewinternet.org/fact-sheet/internet-broadband/> (11 April, 2017).
- Philips, Susan U. 1987. The interaction of social and biological processes in women’s and men’s speech. In Susan U. Philips, Susan Steele & Christine Tanz (eds.), *Language, gender, and sex in comparative perspective*, 1–15. Cambridge: Cambridge University Press.
- Philips, Susan U., Susan Steele & Christine Tanz (eds.). 1987. *Language, gender, and sex in comparative perspective*. Cambridge: Cambridge University Press.
- Pichler, Pia & J. Coates. 2011. *Language and gender: A reader*. Oxford: Blackwell.
- Putnam, Hilary. 1975. The Meaning of “Meaning.” *Language* 7. 131–193.
- Python Software Foundation. 2017. Python 2.7.13 documentation: string — Common string operations. <https://docs.python.org/2/library/string.html> (9 April, 2017).
- Queen, Robin. 2013. Gender, sex, sexuality and sexual identities. In Natalie Schilling-Estes & J. K. Chambers (eds.), *The handbook of language variation and change*, 368–87. 2nd ed. Chichester: John Wiley and Sons.
- Queen, Robin & Julie E. Boland. 2015. I think your going to like me: Exploring the role of errors in email messages on assessments of potential housemates. *Linguistics Vanguard* 1(1). 283–293.
- Rao, Delip, David Yarowsky, Abhishek Shreevats & Manaswi Gupta. 2010. Classifying latent user attributes in twitter. *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, 37–44. ACM. (23 February, 2016).
- Real, Raimundo & Juan M. Vargas. 1996. The probabilistic basis of Jaccard’s index of similarity. *Systematic Biology* 45(3). 380–385.
- Renold, Emma. 2006. They won’t let us play ... unless you’re going out with one of them: girls, boys and Butler’s “heterosexual matrix” in the primary years. *British Journal of Sociology of Education* 27(4). 489–509.
- Rich, Adrienne. 1980. Compulsory heterosexuality and lesbian existence. *Signs: Journal of women in culture and society* 5(4). 631–660.
- Rickford, John R. 1991. Sociolinguistic variation in Cane Walk: A quantitative case study. In Jenny Cheshire (ed.), *English around the world: Sociolinguistic perspectives*, 609–19. New York: Cambridge University Press.

- Rochefort, Charles de, John Davies & Raymond Breton. 1666. *The history of the Caribby-islands*. London: Thomas Dring and John Starkey.
- Romaine, Suzanne. 1982. What is a speech community? In Suzanne Romaine (ed.), *Sociolinguistic variation in speech communities*, 13–24. London: Edward Arnold.
- Romaine, Suzanne. 1996. Review of sociolinguistic theory. *Linguistics* 34(4). 867–70.
- Romaine, Suzanne. 1999. *Communicating gender*. Mahwah: Lawrence Erlbaum.
- Rose, Suzanna & Irene Hanson Frieze. 1993. Young singles' contemporary dating scripts. *Sex Roles* 28(9–10). 499–509.
- Rosenkrantz, Paul, Susan Vogel, Helen Bee, Inge Broverman & Donald M. Broverman. 1968. Sex-role stereotypes and self-concepts in college students. *Journal of consulting and clinical psychology* 32(3). 287.
- Rousseuw, Peter J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20. 53–65.
- Rubin, Donald L. & Kathryn Greene. 1992. Gender-typical style in written language. *Research in the Teaching of English* 26(1). 7–40.
- Rudat, Anja, Jürgen Buder & Friedrich W. Hesse. 2014. Audience design in Twitter: Retweeting behavior between informational value and followers' interests. *Computers in Human Behavior* 35. 132–139.
- Rudder, Christian. 2014. *Dataclysm: Love, sex, race, and identity: What our online lives tell us about our offline selves*. Kindle. New York: Crown.
- Russell, Joan. 1987. Review of Variation in Australian English: The Sociolects of Sydney by Barbara M. Horvath. *Linguistics* 25(2). 434.
- Sallam, A. M. 1980. Phonological variation in educated spoken Arabic: A study of the uvular and related plosive types. *Bulletin of the School of Oriental and African Studies* 43(1). 77–100.
- Sapir, Edward. 1949. Male and female forms of speech in Yana. In D Mandelbaum (ed.), *Selected writings of Edward Sapir in language, culture and personality*, 206–13. Berkeley and Los Angeles: University of California Press.
- Schilling, Natalie. 2011. Language, gender, and sexuality. In Rajend Meshtrie (ed.), *The Cambridge handbook of sociolinguistics*, 218–37. Cambridge: Cambridge University Press.
- Schleef, Erik. 2013. Written surveys and questionnaires in sociolinguistics. In Kirk Hazen & Janet Holmes (eds.), *Research methods in sociolinguistics*, 42–57. Hoboken: Wiley-Blackwell.
- Schler, Jonathan, Moshe Koppel, Shlomo Argamon & James W. Pennebaker. 2006. Effects of Age and Gender on Blogging. *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, vol. 6, 199–205. <http://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-039.pdf> (23 February, 2016).
- scikit-learn developers. 2016. 2.3. Clustering — scikit-learn 0.17.1 documentation. *Clustering*. <http://scikit-learn.org/stable/modules/clustering.html> (21 July, 2016).
- scikit-learn developers. 2017a. TfidfTransformer — scikit-learn 0.18.1 documentation. <http://scikit->

- learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html (15 April, 2017).
- scikit-learn developers. 2017b. API Reference — scikit-learn 0.18.1 documentation. <http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics> (18 January, 2017).
- scikit-learn developers. 2017c. 4.2. Feature extraction: Tf-idf — scikit-learn 0.18.1 documentation. http://scikit-learn.org/stable/modules/feature_extraction.html#tfidf-term-weighting (18 January, 2017).
- Scipy community. 2017. Distance computations (scipy.spatial.distance) — SciPy v0.18.1 Reference Guide. <https://docs.scipy.org/doc/scipy/reference/spatial.distance.html> (18 January, 2017).
- Serewicz, Mary Claire Morr & Elaine Gale. 2008. First-date scripts: Gender roles, context, and relationship. *Sex Roles* 58(3–4). 149–164.
- Shuy, Roger W. 1968. *A study of social dialects in Detroit: Final report*. Washington, D.C.: Office of education.
- Silverstein, Michael. 2003. Indexical order and the dialectics of sociolinguistic life. *Language & Communication* 23(3). 193–229.
- Smith, Christine A. & Shannon Stillman. 2002. What do women want? The effects of gender and sexual orientation on the desirability of physical attributes in the personal ads of women. *Sex Roles* 46(9–10). 337–342.
- Smith, Scott. 2013. Determining sample size: How to ensure you get the correct sample size. *Qualtrics*. <https://www.qualtrics.com/blog/determining-sample-size/> (22 February, 2017).
- Smyth, Ron, Greg Jacobs & Henry Rogers. 2003. Male voices and perceived sexual orientation: An experimental and theoretical approach. *Language in Society* 32(3). 329–350.
- Squires, Lauren. 2012. Whos punctuating what? Sociolinguistic variation in instant messaging. In Alexandra Jaffe, Jannis Androutsopoulos, Mark Sebba & Sally Johnson (eds.), *Orthography as social action: Scripts, spelling, identity and power*, 289–325. (Language and Social Processes 3). Berlin: De Gruyter Mouton.
- Stewart, Mark A., Ellen Bouchard Ryan & Howard Giles. 1985. Accent and social class effects on status and solidarity evaluations. *Personality and Social Psychology Bulletin* 11(1). 98–105.
- Sunderland, Jane. 2006. *Language and gender*. London: Routledge.
- Tagg, Caroline & Philip Seargeant. 2014. Audience design and language choice in the construction and maintenance of translocal communities on social network sites. In Philip Seargeant & Caroline Tagg (eds.), *The language of social media*, 161–185. Palgrave Macmillan UK.
- Tagliamonte, Sali A. 2012. *Variationist sociolinguistics: change, observation, interpretation*. (Language in Society 40). Malden: Wiley-Blackwell.
- Tagliamonte, Sali A. 2016. So sick or so cool? The language of youth on the internet. *Language in Society* 45(1). 1–32.

- Tagliamonte, Sali & Alex D'Arcy. 2004. He's like, she's like: The quotative system in Canadian youth. *Journal of Sociolinguistics* 8(4). 493–514.
- Tannen, Deborah. 1990. *You just don't understand*. London: Virago. (27 November, 2015).
- Thomas, Beth. 1988. Differences of sex and sects: linguistic variation and social networks in a Welsh mining village. In Jennifer Coates & Deborah Cameron (eds.), *Women in their speech communities: New perspectives on language and sex*, 51–60. London: Longman.
- Thomson Reuters. 2015. Web of science. Database. *Web of Science*.
www.webofknowledge.com (11 December, 2015).
- Thorne, Adrian & Justine Coupland. 1998. Articulations of same-sex desire: lesbian and gay male dating advertisements. *Journal of Sociolinguistics* 2(2). 233–257.
- Tillery, Denise. 2005. The plain style in the seventeenth century: Gender and the history of scientific discourse. *Journal of Technical Writing and Communication* 35(3). 273–289.
- Tolman, Deborah L., Renée Spencer, Myra Rosen-Reynoso & Michelle V. Porche. 2003. Sowing the seeds of violence in heterosexual relationships: Early adolescents narrate compulsory heterosexuality. *Journal of Social Issues* 59(1). 159–178.
- Troemel-Ploetz, Senta. 1991. Selling the apolitical. *Discourse & Society* 2(4). 489–502.
- Trousdale, Graeme. 2010. *An introduction to English sociolinguistics*. Edinburgh: Edinburgh University Press.
- Trudgill, Peter. 1972. Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society* 1(2). 179–195.
- Trudgill, Peter. 1974. *The social differentiation of English in Norwich*. Cambridge: Cambridge University Press.
- Trudgill, Peter. 2000. *Sociolinguistics: An introduction to language and society*. 4th ed. London: Penguin UK.
- Uldall, Elizabeth. 1960. Attitudinal meanings conveyed by intonation contours. *Language and Speech* 3(4). 223–234.
- Urban Dictionary. 2017. Urban Dictionary. *Urban Dictionary*.
<http://www.urbandictionary.com/> (18 January, 2017).
- Walker, Abby, Christina García, Yomi Cortés & Kathryn Campbell-Kibler. 2014. Comparing social meanings across listener and speaker groups: The indexical field of Spanish /s/. *Language Variation and Change* 26(2). 169–189.
- Walther, Joseph B. & Kyle P. D'Addario. 2001. The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review* 19(3). 324–347.
- Wardhaugh, Ronald & Janet M. Fuller. 2014. *An introduction to sociolinguistics*. 7th ed. Chichester: Wiley-Blackwell.
- Wenger, Etienne. 1999. *Communities of practice: Learning, meaning, and identity*. Cambridge: Cambridge University Press.
- West, Candace & Don H. Zimmerman. 1983. Small insults: a study of interruptions in cross-sex conversations between unacquainted persons. In Barrie Thorne, Cherie Kramarae & Nancy Henley (eds.), *Language, gender, and society*, 103–19. Rowley: Newbury House.
- Wikipedia. 2014. List of emoticons. (22 March, 2105).

- Winn, Laura L. & Donald L. Rubin. 2001. Enacting gender identity in written discourse responding to gender role bidding in personal ads. *Journal of Language and Social Psychology* 20(4). 393–418.
- Wolfram, Walter A. 1969. *A sociolinguistic description of Detroit Negro speech*. (Urban Language Series 5). Washington, D.C.: Center for Applied Linguistics.
- Zimman, Lal, Jenny Davis & Joshua Raclaw (eds.). 2014. *Queer excursions: Rethorizing binaries in language, gender, and sexuality*. (Studies in Language and Gender). New York: Oxford University Press.