

OMG! L2SPELL ONLINE: THE CREATIVE VOCABULARY OF CYBERLANGUAGE
s(~_^)--b

Laura L. Christopherson

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Information and Library Science.

Chapel Hill
2013

Approved by:

Dr. Connie Eble

Dr. Stephanie W. Haas

Dr. Jeffrey Pomerantz

Dr. Brian Sturm

Dr. Barbara Wildemuth

© 2013
Laura L. Christopherson
ALL RIGHTS RESERVED

Abstract

LAURA L. CHRISTOPHERSON: OMG! l2spell online: The creative vocabulary of cyberlanguage s(~_^)--b (Under the direction of Dr. Stephanie W. Haas)

Increasing use of the Internet has led to a proliferation of online communication and information sharing media. These media, each with its own set of affordances and limitations, are thought to encourage new ways to communicate. Interlocutors refashion general English into abbreviated and often pictographic representations of existing concepts.

Prior research has made suppositions about the effects these media have on communication; for example, that synchronous media (e.g., chat) encourage interlocutors to use more abbreviations (e.g., acronyms) than in asynchronous media (e.g., email). These suppositions, however, have not been fully tested because most studies focus on a single medium. Yet a more comprehensive understanding of this language—hereafter referred to as *cyberlanguage*—as it manifests across various online media is needed as users increasingly employ the Internet for communications. Furthermore, such an understanding may help information professionals improve information tools (e.g., search engines, summarization, surveillance) that currently rely on more standard forms of writing for their success.

The research described here addresses this need by creating and linguistically analyzing a corpus of texts containing 136,529 tokens (23,912 types) that span multiple media (forums, email, text messaging, instant messaging, and chat) and communication situations (business, virtual reference, hobbies, health/well-being). Terms were classified

according to linguistic feature (e.g., acronyms, emoticons). Chi-square tests were used to compare the frequencies of features across media and communication situation.

Contrary to current thinking about “technological determinism,” results show that cyberlanguage feature use varies based on medium and situation, which validates the notion that technology and other situational variables exert influence over communication behavior. New terms are being created all the time online and this suggests rapid language change and linguistic creativity. Interlocutors create new terms to bridge the physical distance between them, such as using surrogate face-to-face cues to make the text seem more like face-to-face speech. However, some cyberlanguage terms and features are quite ordinary and conventional, and may be considered online staples. The number of tokens that contained cyberlanguage features assumed a small portion of the language used online, so fears about cyberlanguage signaling the demise of “proper” English can be allayed.

Acknowledgments

Many people, both in and out of the doctoral program, supported me through this process, and I would like to extend my deepest gratitude for all they did for me.

- Stephanie Haas, my advisor and mentor, for shepherding me through this process,
- Barbara Wildemuth, for advising and listening to me about methodological concerns,
- Brian Sturm, for keeping me grounded, making me laugh, and sharing wonderful insights,
- Connie Eble, for teaching me more about linguistics and slang, and sharing a love of “deviant” language,
- Jeff Pomerantz, for showing support, providing the NCknows conversations, and questioning me,
- Clifton Barnett, for programming most of the scripts for me and being so generous with his time,
- Joshua Bert Purvis, for being my most excellent outside coder and “unknown” term disambiguator,
- David R. Jackson, for writing a data collection script,
- Eliah Hecht and Shayne Muelling for inspiring me to write about gaming language,
- Pam Sessoms and the UNC University Libraries for providing IM conversations,
- Susana Sotillo, for sharing her SMS corpus with me,
- Caleb Tucker-Raymond for sharing the L-net conversations with me,

- Craig Martell, for sharing the NPS corpus with me,
- AOL for granting permission to use their chat conversations (AOL is a registered trademark of AOL LLC),
- Dan Piraro, for the use of his fabulous OMG! Cartoon in the Preface,
- Elizabeth A. Evans (Libby), for inspiring me and being a sounding board,
- Mike Brown, for helping me improve my writing,
- Angela Murillo and Sarah Ramdeen, for picking me during a dark time,
- Nancy Honeywell, for being an awesome mentor,
- My RTP LB family: thank you for teaching me so much, making me laugh, helping me to remember not to take myself so seriously, holding me up, and walking with me through everything

Preface



Cartoon courtesy of Dan Piraro

The title of this dissertation can be translated as follows: *Oh my god! Learn to spell online: The creative vocabulary of cyberlanguage*. The emoticon at the end is a winking face using a tilde for the wink, and giving a thumbs up, using the --b for the arm and thumb on the hand.

Table of Contents

LIST OF TABLES	xiv
LIST OF FIGURES	xvi
Chapter	
I. INTRODUCTION.....	1
II. LITERATURE REVIEW	8
Introduction	8
Cybermedia.....	9
Media characteristics.....	12
Synchronicity.....	12
Participant scale.....	14
Message persistence	17
Privacy.....	19
Anonymity / pseudonymity	20
Message length	22
Compositional and viewing ease.....	23
Quoting and linking.....	24
Conclusion.....	25
Cybermedia descriptions.....	26
Forums.....	26
Email.....	28
SMS	32

IM	35
Chat.....	38
Conclusion.....	42
Theories about Communication in Lean Media	43
Social presence theory	43
Media richness	45
Lack of social context clues	46
Channel expansion theory.....	47
Social information processing.....	47
Conclusion	49
Other Situational Variables: Genre.....	49
Purpose.....	50
Interlocutors	51
Norms, expectations, and the situation	51
Content.....	52
Form / structure.....	52
Conclusion	53
Cyberlanguage and Its Characteristics	54
Abbreviations.....	58
Acronyms / initialisms.....	59
Shortenings.....	59
Clippings.....	60
Single-letter forms.....	61
Letter homophones	61
Number homophones.....	62
Symbolic substitution	62

Conjunctions and disjunctions.....	63
Punctuation omission.....	63
Non-standard use of lowercase.....	64
Surrogate prosodic cues	64
Onomatopoeic expression	65
Phonetic respellings.....	65
Offsetting punctuation	66
All caps.....	66
Letter duplication.....	67
Punctuation duplication.....	67
Surrogate proxemic cues	67
Emoticons	68
Emotes	68
Pointing.....	69
Pictograms	69
Other features	70
Misspellings and typos	70
Repairs.....	71
Addressivity.....	71
Reduplication.....	72
Other word-creation processes	72
Conclusion	73
Word Creation.....	77
Word-formation	78
Productivity and creativity.....	80
Language play.....	83
Conclusion	84

Conclusion	86
III. Methods	91
Introduction	91
Corpus Design	94
Balancing criteria	97
Corpus size.....	102
Corpus Creation	105
Forums	111
Email lists.....	118
SMS.....	122
IM.....	123
Chat	130
Corpus cleaning.....	135
Conclusion	137
Analysis	142
Classification.....	145
Inter-coder reliability test.....	147
Chi-square tests	152
Conclusion and Limitations	156
IV. Findings	160
Introduction	160
Synchronicity and Persistence	165
Participant Scale	170
Anonymity	174
Message Length	176

Compositional Ease	179
Viewing Ease	184
Topic	186
Purpose	193
Conclusion	199
V. Discussion	206
Major Findings	206
Technological determinism.....	206
Ordinariness and conventionality.....	210
“Proper” English	211
Conclusion	213
Specific Features	214
Acronyms / Initialisms.....	214
Symbolic Substitution	218
Lowercase and All Caps	222
Onomatopoeic Expression	224
Phonetic Respellings.....	226
Offsetting Punctuation	228
Emoticons.....	231
Emotes.....	235
Repairs, Addressivity, and Compounds / Space Omission.....	242
Additional Examples of Creativity	243
Conclusion	246
VI. Conclusion	248
APPENDICES	256

Appendix A: A Note about the Term <i>Cyberlanguage</i>	256
Appendix B: Differences between Speech and Writing.....	260
Appendix C: Support for Topic and Purpose Classifications	261
Appendix D: Coding Rules.....	271
Appendix E: Table of Signs and Symbols	285
REFERENCES.....	287

List of Tables

1. Cybermedia and their characteristics in the most typical scenarios.	42
2. Cyberlanguage features, definitions, examples, and sources.....	73
3. Dimensions of cyberlanguage corpora.....	103
4. Word counts for the forums section of the corpus.....	118
5. Word counts for the email lists section of the corpus.....	121
6. Word counts for the SMS section of the corpus.....	123
7. UNC Ask a Librarian sampling statistics.....	126
8. NCKnows sampling statistics.....	127
9. L-Net sampling statistics.....	128
10. Word counts for the IM section of the corpus.....	129
11. Word counts for the chat section of the corpus.....	134
12. Corpus details.....	139
13. Word counts for media sections of the corpus.....	141
14. Word counts for topics and purposes.....	141
15. Inter-coder telability kappas.....	151
16. Media factor comparisons.....	155
17. Genre factor comparisons.....	156
18. Counts for words collected, general/standard English, and cyberlanguage terms....	160
19. Feature frequency and percent.....	161
20. Comparison of features among the five media.....	163
21. Comparison of features between synchronous and asynchronous media.....	166
22. Comparison of features by participant scale.....	171

23. Comparison of features by the degree of anonymity afforded by the medium	174
24. Comparison of features by message length restrictions.....	177
25. Comparison of features by compositional ease	180
26. Comparison of features by viewing ease	185
27. Comparison of features by gaming, technology, gaming technology, and other topics.....	187
28. Comparison of features by gaming, technology/gaming technology, and other topics.....	190
29. Comparison of features by serious, recreational/leisure-oriented, mixed, and ambiguous purposes.....	194
30. Comparison of features by serious and recreational/leisure-oriented purposes.....	196
31. Comparison of features by non-recreational (serious and ambiguous) and recreational/leisure-oriented purposes	198
32. Features that are common to the five media, to the three core topics, and the two core purposes; “x” signifies features with insignificant chi-square values.....	200
33. How features vary in different online communication situations; “x” signifies higher than expected frequency for statistically significant comparisons	202
34. Examples of creative word-creation	244
35. Frequency and proportion of types of features	253

List of Figures

1. One-to-one structure	15
2. One-to-many structure	15
3. Many-to-many structure.....	16
4. A hypothetical line of chat.....	135
5. Punctuation and numerals keypad on an iPhone.....	183

Introduction

The Internet has become an “embedded” part of people’s everyday lives (Haythornthwaite & Wellman, 2002, p. 6). People spend time emailing, texting, chatting, shopping, reading news, seeking health information, participating in auctions, planning travel, looking for love, playing games, and more. The Internet is used for all sorts of purposes: work, education, research, and recreation.

Wellman (2004, p. 23) says that the Internet has “burrowed” into his life; it is not separate from the rest of his life. He explains that “the longer people have been on the Internet, the more they use it” (p. 26). In their 2005 report on Internet use, the Pew Research Center (Pew Research Center, 2005, p. 58) reports that “70 million American adults logged onto the internet”—a “37% increase from the 52 million adults who were online on an average day in 2000.” In a 2012 Pew survey, 85% of adults surveyed used the Internet on a variety of computerized devices including mobile devices (Rainie, 2012). Rideout, Foehr, and Roberts (2010) explain that in 2009, 93% of 8-18 year-olds lived in a home with a computer (a 20% increase from 1999), 84% lived in a home with Internet access (a 37% increase from 1999), and 66% owned a cell phone as opposed to 39% in 2004. “And, because of media multitasking, the amount of media content consumed during that period has increased from 7½ hours [per day] in 1999 to 8½ hours in 2004 and to more than 10½ hours in 2009” (Rideout, Foehr, & Roberts, 2010, p. 11). The need to “keep up” motivates Internet use (Haythornthwaite & Wellman, 2002, p. 10).

Arenas for communicating and connecting with many others—such as MUDs and MOOs,¹ email, chat rooms, discussion forums, instant messaging, text messaging, Twitter, and massively multiplayer online games (MMOGs)—were developed for the Internet. These venues do not “replace more traditional offline forms of contact” but rather complement them, “increasing the overall volume of contact” (Wellman, 2004, p. 25). They help to bring people together, especially people who would not ordinarily mix. The Internet “extends communities in the real world” (Wellman, 2004, p. 22), and is an avenue for forming new partnerships, collaborations, and friendships that might not otherwise be possible. “Rather than functioning in discrete, bounded groups—at home, in the community, at work, in organizations—people move as individuals between various fuzzily-bounded networks” (Haythornthwaite & Wellman, 2002, p. 10). This expansion of the social, personal, and professional aspects of people’s lives through use of these online media has altered the way people relate to information, and has fostered new forms of communication.

Users began to refashion general (or standard) English into “abbreviated and sometimes pictographic representations of existing concepts where layers of meaning are packed into a few keystrokes” (Christopherson, 2013, Online Communication section, para 1). Documented by many researchers—such as Baron (2003, 2008, 2010), Cherny (1999), Crystal (2006, 2008), Danet (2001), Hård af Segerstad (2002), Herring (2001, 2002, 2012), Lewin and Donner (2002), Werry (1996)—this online language, referred to in this dissertation as *cyberlanguage* includes abbreviations of all kinds and surrogate face-to-face cues—textual substitutes for proxemic and prosodic cues that are missing in most online communication settings. (See Appendix A: A Note About the Term Cyberlanguage.)

¹ *MUD* stands for *multiple user dimension* or *dungeon*. *MOO* stands for *MUD object oriented*. These were early

Cyberlanguage is a result of user response and adaptation to the constraints and affordances imposed by these new media and other aspects of the online communication situation.

Online, conversational media—or *cybermedia*—possess certain characteristics, certain limitations and features, which seem to constrain or empower interlocutors in their communication activities. Interlocutors adapt by using word-creation processes in new ways to create new words, phrases, and syntax. For example, some cybermedia—such as chat, IM, and text messaging—restrict messages to a certain character length. It is believed that these restrictions encourage interlocutors to use more abbreviations (Herring, 2007). Thus acronyms such as *lol* for *laughing out loud*, letter homophones such as *u* for *you*, and shortenings such as *prolly* for *probably* may be more prevalent in media with message length restrictions. Cybermedia that allow multiple interlocutors to converse simultaneously have been shown to exhibit very playful language. For example, Cherny (1999) and Werry (1996) found many examples of playful coinage in their examinations of chat—a many-to-many, synchronous medium.

Similarly, genre—social and contextual aspects of the communication situation such as topic of discussion and purpose for communicating—have been thought to also influence language production (Herring, 2002). Hård af Segerstad (2002, p. 199) explains that the “linguistic characteristics” of messages are not solely determined by the technical aspects of the medium, but are also attributable to interlocutors’ “interpersonal relationships and their reasons for communicating.” For example, Herring (2001) noticed more contractions in conversations centered on “fun” topics.

“What is truly remarkable is that so many people have learned so quickly to adapt their language to meet the demands of the new situations, and to exploit the potential of the

new medium so creatively to form new areas of expression” (Crystal, 2006, p. 276). Internet communication is evidence of rapid language change. The data used in this dissertation study, that provides conversation from five cybermedia and several different topics and purposes, contains thousands of terms that are not found in general/standard English dictionaries and therefore may be new to the reader of this dissertation.

Crystal (2006, 2008b) considers this online language to be evidence of linguistic creativity. It is an example of how technology shapes communication and information behavior, and how people adapt to and capitalize on changes in the environment. Crystal (2006, p. 272) believes the Internet “is going to ‘change the way we think’ about language in a fundamental way.” Cyberlanguage may be a “development of millennial significance” (Crystal, 2006, p. 272), “an expansive new linguistic renaissance” (Tagliamonte & Denis, 2008). As more people use the Internet and become more comfortable with its affordances and limitations, more new terms may be coined, and they may be used outside of online contexts. In April of 2011, the BBC reported the addition of *LOL* (*laughing out loud*, a popular acronym used online) to the Oxford English dictionary (Morgan, 2011).

In addition to being a shining example of rapid language change, cyberlanguage has implications for information seeking, capture, and use. As cyberlanguage becomes more widely used, information professionals will need to respond by rethinking the design of techniques and tools used for the purposes of searching, capturing, organizing, and monitoring information. Currently there is no lexicon of cyberlanguage that information professionals may refer to when attempting to disambiguate messages. So information retrieval and surveillance tools, for example, may be limited to best guesses about message meaning. Information tools must become as facile with cyberlanguage as they are with

standard language so that information retrieval, summarization, document clustering, and other information tools can provide users with the best possible search and use experiences.

As Sager (1990, p. 7) explains,

In general, a greater understanding of the paradigmatic units of special subject languages is of considerable advantage to information science. The practical objectives of terminology, i.e. to achieve greater unity, consistency and clarity of expression in special communication would greatly simplify the work of information scientists.

Furthermore, online, conversational media are not just for recreational or frivolous purposes. Child pornography rings have been shown to operate in chat rooms (CNN.com, 2006) and Iraqi death squads have used them to entrap victims (PinkNews, 2007). Brachman (2006, p. 150) explains that the U.S. government monitors jihadi communication in “email, chat rooms, online magazines, cell phone videos, CD-ROMs, and even video games for immediate intelligence indicators and warnings.” If language used in the performance of such underground activities is not understood, it becomes difficult to take actions to prevent or thwart such activities. Thus, an understanding of cyberlanguage may also help information professionals develop better intelligence surveillance tools.

The study described here aims to provide a detailed description of cyberlanguage and its use. The researcher conducted a review of the research into cyberlanguage and compiled a list of the linguistic features identified by other researchers. She created a corpus of texts from five online conversational media—specifically chat, IM, text messaging, email, and forums—spanning several topics and purposes; and she examined terms in the corpus for the presence of the features found during the literature review. The researcher compared feature

frequencies using chi-square tests to determine what features differ across media, topics, and purposes, and what features are common across them. The goal was to test assertions made by other researchers about the influences cybermedia and other situational variables may have on language production.

At the time of his writing, Crystal (2006) explains that a systematic, empirical observation of this sort has yet to be pursued and that no corpus, of the sort described here, has been created. To the researcher's knowledge, the research reported here is the first to do so. It is possible that the length of time required for such an analysis has been a deterrent for other researchers. Most cyberlanguage research has focused on a single medium in isolation. Therefore these studies could not make comparisons across media to provide a broader, more comprehensive view of cyberlanguage or to verify assertions made about its use in different media and different genre situations, although some studies have attempted comparison of language across media via ex post facto analysis of findings from multiple studies. For example, Baron (2008) compared findings from two studies of SMS and IM, and Hård af Segerstad (2002) compared findings from separate studies of email, SMS, IM, and chat.

The study described here distinguishes itself from prior research in that its use of a large corpus that spans multiple media types, topics, and purposes, and is framed by a consistent set of research questions and a consistent methodological approach. The results provide a broader description of cyberlanguage and how features may vary (or not) depending upon the communication situation. As such, these results should provide a clearer picture of how technology may influence users and how users may creatively adapt their behavior to suit technological change. In addition to the corpus, specific products resulting from this research may include a lexicon of cyberlanguage terms and their usage, and a list of

rules for automatically detecting cyberlanguage terms in text samples. These tools could be used by information professionals to improve information retrieval, summarization, surveillance, and other information seeking and use processes.

The rest of this document is organized as follows. A review of the literature describes and discusses (a) cybermedia and their characteristics as possible influencing variables, (b) theories posed to describe cybermedia's influence on online communication, (c) aspects of genre and how these might influence language production, (d) cyberlanguage features as identified by other researchers, and (e) word creation including word-formation, productivity, creativity, and language play. The methods for creating the corpus, classifying terms by feature, and comparing frequencies using chi-square tests follow. Research findings are then detailed and specific features are discussed more fully in the Discussion section. Examples of terms that exhibit certain features and demonstrate linguistic creativity are also provided.

Literature Review

Introduction

Cyberlanguage—used in online media such as chat, instant messaging, text messaging, games, forums, and other social media—is characterized by the refashioning of standard English into abbreviated and often pictographic representations of existing concepts where layers of meaning are packed into a few keystrokes. It is a result of user response and adaptation to the constraints and affordances imposed by these media, such as character length restrictions or small screen/window size. Cyberlanguage demonstrates human understanding of the principles of word-creation and language production, and human capacity for creatively exploiting those principles to reshape language to suit the communication situation (Crystal, 2006, 2008b).

This literature review describes earlier research on the subject of online communication. It begins by describing online, conversational/social media—hereafter referred to as cybermedia—and the characteristics of these media that may influence the use of certain linguistic features in word creation. Following this will be a discussion of theories applied to the study of cybermedia to frame discussions about the degree to which these media permit rich and intimate conversations.

In addition to considering the medium's influence on language production, other aspects of the communication situation—such as topic of discussion and purpose for

communicating—and their potential influence on language production will be discussed. These aspects are discussed as facets of genre.

Then a description of online communication and a catalogue of linguistic features of cyberlanguage, as documented by earlier researchers, will be discussed in detail. Additional linguistic concepts, such as word-formation, productivity, linguistic creativity, and linguistic play, will be defined because it is believed, by this researcher and scholars such as Crystal (2006, 2008b), that cyberlanguage is evidence of new and innovative word-creation strategies that demonstrate the creativity of cyberusers. Additionally, examples of creativity will be sought in the study corpus and so concepts around this topic must be defined.

In sum, this literature review aims to provide an overview of prior research and thinking on the subject of online language—or cyberlanguage—including the different communication situations where it may be observed, its characteristics and features, and how interlocutors may exploit the rules of language to create cyberlanguage. The purpose of this literature review is to set the stage for the analysis described in later sections of this document, in particular, for testing assertions about cyberlanguage that have hitherto not been tested.

Cybermedia

Biber's (1988) definition for *channel* and Halliday's (2007b) definition for *mode* are relatively similar. They largely refer to classifying language as either speech or writing, but may also include other communication forms, such as drum systems (Biber, 1988) or signs, as in sign language or Braille (Zawada, 2005, p. 70). In this paper, the term *medium* will be used instead and will be defined as the “conduit,” “pipeline,” “avenue,” “venue,” or “arena”

for communication. Different language varieties, different genres, and different channels such as speech, writing, and graphics are conducted through different media. For example, a chat program is a medium that allows interlocutors to communicate with written language and sometimes with graphical symbols depending upon the affordances of that particular chat program. A telephone is a medium that allows interlocutors to communicate via speech in a variety of languages. One can also consider two or more people in close physical proximity to one another—i.e., face-to-face situations—as a particular medium for communication.

Interlocutors “develop new ways of using language in the process of communication in new media, and these new discourse practices are sometimes tied to constraints and possibilities afforded by the media themselves” (Johnstone, 2008, p. 196). “Factors such as screen size, average typing speed, minimal response times, competition for attention, channel population and pace of channel conversations all contribute to the emergence of certain characteristic properties” (Werry, 1996, p. 53). Johnstone (2008) explains that certain media can be better for certain purposes over other media, can require different ways to interpret or recall information, can afford different types of activities (e.g., collaboration vs. monologue), can set different interpersonal tones, and can encourage people to regulate their behavior differently. Because language and the level of interactivity are so sensitive to medium, online communication becomes “far more complex and variable than envisioned by early description” (Herring, 2001, p. 613).

This paper focuses on media that are accessible through the Internet by using either a computer or hand-held² computerized device, and permit conversation between two or more interlocutors. Conversation is defined as a dynamic, back-and-forth flow of comments and

² Hand-held devices can include iPod Touches, cellphones with messaging capability, smartphones, or personal digital assistants (PDAs) for example.

responses—a repartee, a discussion—by two or more interlocutors where thoughts tend to be shared in an extemporaneous manner with less planning and editing than one might see in more formal texts, such as scholarly articles, brochures, news articles, novels, etc. The text is necessarily dialogic, not monologic. In other words, the text is not for the purposes of one-sided broadcasting of thoughts and ideas. Conversation emerges, unfolding organically over a period of time and having an unpredictable focus, rather than being created holistically with a predetermined focus as one might find in scholarly articles, news articles, novels, etc.

Therefore, static webpages and blogs will not be discussed because they may not invite user feedback, and are thus primarily monologic. Of webpages, Crystal (2006, p. 206) concludes, “if we are looking for Internet distinctiveness, novelty, and idiosyncrasy – or wishing to find fuel for a theory of impending linguistic doom – we are not likely to find it here.” Twitter and Facebook wall posts will also not be discussed because they are primarily used for one-off broadcasting of thoughts. These media, although sometimes encouraging conversation, by and large do not result in conversation that satisfies the above definition. Instead, discussion forums, email (which also includes email lists such as listservs), SMS (short message service—text messaging on mobile devices), IM (instant messaging), and chat (including gaming chat) will be explored in this literature review. Conversations taken from these media will be analyzed in this dissertation study.

What follows is a discussion of media characteristics that are thought to have an influence on language production in online communication. The following characteristics—many which were drawn from Herring’s (2007) faceted classification scheme—will be described in full:

- synchronicity,

- participant scale (one-to-one, one-to-many, many-to-many),
- the degree of persistence of message,
- the degree of privacy of conversations,
- user anonymity/pseudonymity,
- message length restrictions,
- compositional and viewing ease ,
- support for quoting or linking to other messages.

Then each of the media used for analysis in this paper—forums, email, SMS, IM, and chat—will be further described. The section will conclude with a table outlining the characteristics of each medium.

Media characteristics

Synchronicity

Some media such as IM and chat offer synchronous engagement where interlocutors can converse simultaneously. Other media, such as SMS, email, and forums allow interlocutors to respond at different points in time. Synchronous media are thought to lead to brevity, playful and phatic speech, typographic and orthographic innovations, disrupted turn-taking, less structural complexity, and an informal style (Herring, 2002). Dresner (2005, Visual Spatiality and Textual Chat section, para. 4) comments that synchronicity “exercise[s] a powerful influence over structural complexity.” Users “under the pressure to type at a conversational pace” may sacrifice linguistic complexity (Herring, 2002, p. 139).

Just as the structure of unplanned speech reflects cognitive constraints on real time language encoding, for example in length of information units, lexical density and degree of syntactic integration, so too synchronous modes of CMD³ impose temporal constraints on users that result in a reduction of linguistic complexity relative to asynchronous modes. (Dresner, 2005, Visual Spatiality and Textual Chat section, para. 4)

Dresner (2005) suggests that the immediacy and speed of communication often sacrifice complexity. More abbreviated forms may appear in synchronous communication because, after all, “abbreviations speed things up” (Crystal, 2008b, p. 65). Dresner (2005) indicates that in synchronous environments, editing is simply not practical. Users cannot keep up with the unceasing flow of information if they take the time to carefully reflect on their assertions and edit any mistakes (Crystal, 2006). The tide of the conversation will sweep them away. Ferrara, Brunner, and Whittmore (1991) explain that there is a willingness in online discourse to allow others to see each other’s mistakes, and that this is unique to these communication situations. The “first draft quality” of such discourse is an accepted convention (Ferrara et al., 1991, p. 25); and so it is not seen as odd or irresponsible, much in the same way that performance errors in speech are viewed as part and parcel of normal communication.

Synchronous modes appear to be better suited for social interaction and contribute more to social presence; whereas asynchronous modes tend to be better for “more complex discussions and problem solving” and are often found to be more linguistically complex (Herring, 2002, p. 135). In synchronous modes, users opt for brief exchanges that save time and keystrokes (Ferrara et al., 1991). Synchronous communication “reads like and to a certain extent acts like conversation” (Davis & Brewer, 1997, p. 2).

³ *CMD* stands for *computer-mediated discourse*.

Herring (2007, Situation Factors section, para. 11) notes that “all other things being equal, for example, synchronous CMD is more likely to be informal in register and playful in tone than asynchronous.” Crystal (2006, p. 135) explains that “it is the synchronous interactions which cause most radical linguistic innovation.”

Asynchronous media allow interlocutors more time for planning, editing, and managing self-presentation, so messages may be more complex and formal (Herring, 2002). Interlocutors may be less likely to be caught up in the moment and act uninhibitedly. Asynchronous conversations are less interactive because of the delay (Davis & Brewer, 1997). Danet (2001) concludes that interactivity is an important contributor to a playful tone. Synchronous communication is more analogous to a theatrical performance where, although scripted, blocked, and directed, it is open to the unexpected. Asynchronous correlates more to a movie where mistakes are fixed and takes are done over and over until it is perfect; there is substantially less room for the unpredictable than can occur in live theater.

Participant scale

Scale refers to how many participants may converse at one time. One-to-one (1:1) communication is characterized by one interlocutor commenting or responding to one other interlocutor at a time. Many media (and sources⁴) may offer opportunities for conversation at multiple scales (e.g., World of Warcraft⁵ chat) or may foster conversation that changes scale over time (e.g., forums). World of Warcraft (WoW) chat offers chat channels that not only

⁴ *Source* refers to a specific instantiation of a medium. For example, chat is a medium. Sources for the chat medium include AOL chat, World of Warcraft chat, etc. A discussion forum is a medium, and discussion forums can be offered by a variety of sources: AOL, World of Warcraft, EverQuest, Teenspot.com, etc.

⁵ World of Warcraft is an *Massively Multiplayer Online Game (MMOG)*, which is a type of virtual world game. World of Warcraft is also referred to as *WoW*, (<http://www.worldofwarcraft.com>).

allow multiple individuals to converse at the same time (N:N), but also allow two individuals to have a private conversation (1:1). Although a forum posting may be initiated as a 1:N broadcast, it can result in a N:N discussion.

IM, chat, email, and SMS afford one-to-one (1:1) communication.

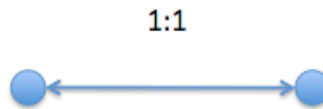


Figure 1: One-to-one structure.

One-to-many (1:N) communication is characterized by one person sending a message to many recipients at a point in time, almost like broadcasting to an audience. Forums and email lists are the main types of media exhibiting 1:N.

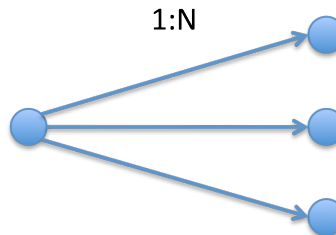


Figure 2: One-to-many structure.

Many-to-many (N:N) communication is characterized by multiple interlocutors carrying on a conversation simultaneously. Individual interlocutors may respond to all members of the group. Chat is the best example of N:N.

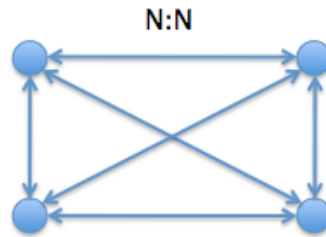


Figure 3: Many-to-many structure.

Conversation within these media—particularly those that support N:N communication where several interlocutors may be talking at once—is sometimes “claimed to be interactionally incoherent” (Herring, 2001, p. 618) and can require greater conversational management on the part of interlocutors. Incoherent conversation is disjointed and subject to disruption and topic breakdown. Herring (2001, p. 618) explains that “disrupted turn *adjacency* [is] caused by the fact that messages are posted in the order received by the system, without regard for what they are responding to.”

Turn-taking is so fundamental to conversation that most people are not conscious of its significance as a means of enabling interactions to be successful. But it is a conversational fact of life that people follow the routine of taking turns, when they talk, and avoid talking at once or interrupting each other randomly or excessively. Moreover, they expect certain ‘adjacency-pairs’ to take place: questions to be followed by answers, and not the other way around; similarly a piece of information to be followed by an acknowledgement, or a complaint to be followed by an excuse or apology. (Crystal, 2006, pp. 35-36)

When several people’s comments are entered seemingly at once in a N:N situation, they may appear in the chat window in the order in which the chat program received them but that order may not reflect a logical sequence of utterances. For example, an answer to a question may come in several lines after a second question has been posed, thus making it difficult to

determine which question the answer is addressing. Dresner (2005) speculates that it is possible that this lack of coherence may serve a purpose; it may make a chatroom, for example, seem more playful and light-hearted. Because topics swell up and then decay, because threads weave through conversations and then are broken, because the conversation rises and falls and turns like a roller-coaster, the conversation can seem ebullient and vivacious, bringing people together in a party-like atmosphere. Being adept at negotiating seemingly chaotic situations by demonstrating facility with the language affirms group identity.

It would seem that, when social advantages are so great, people make enormous semantic allowances. Several authors make the point that the presence of linguistic confusion and incoherence could be inherently attractive, because the social and personal gains – of participating in an anonymous, dynamic, transient, experimental, unpredictable world – are so great. (Crystal, 2006, p. 175).

Message persistence

Persistence can be viewed in two respects: message visibility and message re-use. Message visibility refers to how long the message appears on the screen before scrolling out of the buffer⁶ and being concealed from interlocutors. Message re-use refers to whether the message is stored in some accessible online location, available for re-use and reviewing by interlocutors.

IM and chat scroll quickly and sometimes have limited buffer sizes; so at a certain point, the source medium may cause messages to scroll out of the viewing area, making it impossible for interlocutors to see prior messages. This makes it more akin to speech than if

⁶ A buffer is a region of computer memory where data can be stored temporarily. Some chat applications, for example, will only retain so much text before it has to push text out to make room for more.

messages lingered longer. In face-to-face speech, interlocutors are not able to observe their words after they have been produced (Hård af Segerstad, 2002). “Small buffers also increase the likelihood that language will be structurally abbreviated” (Herring, 2007, Medium Factors section, para. 8). SMS, email, and forums afford the user more control over visibility. The messages of these media do not tend to disappear after some capacity limit (i.e., buffer size) has been reached.

With SMS and email, messages are often automatically retained unless users choose to specifically delete the message. Thus they afford more message re-use than IM and chat where sources may not automatically retain messages. Instead with chat or IM, users may choose to log or copy messages themselves. Even when a message (of whatever kind) is removed from a personal computing device (whether by the source medium or by the user himself), it is still possible the system administrator for that particular source may elect to keep a log of messages on a server. For instance, with IMAP⁷ email protocol, messages are downloaded to a user’s computer (or computerized device) but copies are also retained on the server. So even when a user deletes a message from his computer, a copy might still be stored and retrieved from the mail server. The greater the impermanence, the more likely one might notice greater interactivity (Herring, 2002), “more urgency and energetic force” (Crystal, 2006), with interlocutors vying for attention (Werry, 1996). Such ephemeral conversations with less time for editing and review may be less predictable (Crystal, 2006) and may seem more spontaneous like speech (Hård af Segerstad, 2002).

⁷ *IMAP* stands for *Internet Message Access Protocol*. *POP* stands for *Post Office Protocol*. Both are protocols or means by which a client machine may access email messages housed on a mail server. The fundamental difference between these two is that messages are retained on the email server with IMAP but not necessarily with POP. In a way, IMAP provides a window into the server, while POP downloads all email from the server to one’s computer (and so copies are not stored on the email server once copied to the client machine).

Privacy

Media have varying degrees of privacy, with privacy being defined as limiting the viewing of messages to select individuals as opposed to anyone having access to messages. In a way, this definition of privacy can be viewed as pertaining to exclusivity and the level of intimacy of conversations. For instance, email lists may be open to the public (viewable to anyone who locates a webpage that lists and displays messages) or closed (only viewable by those subscribed). In AOL (formerly America Online) chat rooms, messages are viewable by anyone who opens the webpage to view the chat room. The person viewing does not have to be logged into and participating in the chat room in order to view the messages. However, other chat rooms may require logging in to see messages. Email privacy is more variable. Although a single individual sends a message to another single individual, this doesn't preclude the recipient from forwarding the email to someone else. Private chats in World of Warcraft, although seemingly private to the two participating individuals, are probably accessible by game administrators if such conversations are logged on the game server. So the degree of privacy of a particular medium is not solely a function of the medium itself (i.e., not all sources providing chat rooms have the same affordances). Instead the degree of privacy is a function of the specific implementation of the medium by a particular source (e.g., AOL chat vs. WoW chat) and how interlocutors choose to use it.

Danet (2001, p. 144) uses the term “participatory spectacle” to suggest that performance and play may be more likely to occur in public, N:N conversations, such as chat. The more one is on the stage, so to speak, perhaps the more likely s/he is to act in a less inhibited manner. Crystal (2006) suggests that some chat rooms are akin to cocktail parties, with everyone talking at once. So much chatter among so many individuals may indeed lead

to a participatory spectacle. In Cherny's (1999) analysis of a MOO (a chat room of sorts), word play was ubiquitous.

Anonymity / pseudonymity

Race, sex, socioeconomic status, and the like are often obscured online because interlocutors communicate at a distance and pseudonyms are often used in place of real names (Androutsopoulos 2006; Herring, 2001). Some media may allow interlocutors the opportunity to conceal their identities by selecting userids (also known as *handles*, *pseudonyms*, and *nicknames*) that do not resemble their real name. Although this may not result in total anonymity, it affords some degree of anonymity that may or may not last depending on a number of things, such as a decision to disclose one's identity or leaking identifying information about others.

The anonymity of many of these media allows users to alter their self presentation in a way that is not possible in face-to-face communication (Reid, 1991). Not being bound by race, gender, age, and the like, users can present a customized representation of self, thus experimenting with identity (Reid, 1991). Interlocutors can "hide behind the text" (Hård af Segerstad, 2002, pp. 128). "Operating behind a false persona seems to make people less inhibited: they may feel emboldened to talk more and in different ways from their real-world linguistic repertoire" (Crystal, 2006, p. 54). Anonymity has been found to lead to increased self-disclosure, play, and flaming (Herring, 2007).

The context within which the medium is used often influences the selection of pseudonyms. For instance, in WoW the vast majority of pseudonyms this researcher has noticed do not resemble real names. Instead, WoW players often choose fantasy names,

clever turns of phrases, and jokes. For instance, one player this researcher has interacted with uses “The Lord of the Rings” terminology for his pseudonyms. Another player creates names that are clever turns of phrases and puns, such as *Shamuljckson* and *Bubblefett*. *Shamuljckson* is a play on the actor’s name but also on the class⁸ of the character being played, a *shaman*. Similarly, *Bubblefett* is a play on the name of the character, Boba Fett, in the “Star Wars” trilogy and on the class played—a paladin that has a special ability of being able to create a protective bubble around himself. *Frostitute*—part joke and part word play—is a name selected by a player whose character, a sexy female mage, uses frost spells.

Many email accounts are often assigned at places of work, thus userids resembling real names may be more common than chat ids. So whether the medium will be used for work or leisure and how much control the user has over userid selection can have an impact on the creation of userids. Furthermore, with email addresses, the domain name that appears after the @ symbol can reveal additional information about the person, such as which organization s/he works for. With SMS, usually a phone number serves as the handle. Email, IM, and SMS interlocutors may have greater familiarity with each other prior to establishing communication because to be able to send a message often means that at some point in the past there has been an exchange of contact information (including names). However, in forums, email lists, and chat, communication can be established without knowing any contact information for the recipient(s). It is believed that the greater the degree of anonymity of a medium, the lower the social accountability and therefore the higher the risk for antisocial behavior to occur (Herring, 2002). If no one truly knows who the culprit is, then imposing sanctions for inappropriate behavior becomes difficult if not impossible.

⁸ Every player in WoW may choose to play a character of a particular race and class. Races include things like elves, orcs, humans, dwarves, etc. Classes are essentially roles that the character fulfills such as mage, warlock, shaman, druid, paladin, warrior, etc.

Message length

These media place different restrictions on message length. Email and forums tend to allow messages of any length. IM, SMS, and chat sources tend to restrict messages to a certain character length. For example, WoW chat restricts messages to 255 characters (Collister, 2008), and SMS to 160 (Crystal, 2008b). Herring (2007) indicates that the smaller the buffer size and allowable message length, the more likely the communication is to be abbreviated and less reminiscent of prose. With a limited number of characters to work with, interlocutors may need to be more creative in abbreviating messages so as to pack in as much information as needed to get one's point across succinctly. Limited message lengths coupled with synchronous, N:N situations may lead to even more abbreviations because interlocutors are faced with multiple pressures to write quickly and succinctly: limited space within which to compose a message, immediacy, and many interlocutors at once competing for the floor.

Message length restrictions may also contribute to difficulty in achieving interactional coherence. These limits may force interlocutors to use multiple lines to complete a turn, which then makes that turn more susceptible to interruption. Collister (2008), in her conversational analysis of WoW chat, indicates that interlocutors do this by contributing increments of their messages, one transmission at a time, gluing on additional information to the earlier transmission. However, some users will end chat lines with ellipses to indicate that the thought is yet to be completed, thus providing means by which to connect messages into a more coherent turn.

Compositional and viewing ease

As with message length restrictions, it has been suggested that the more difficult it is to compose and view messages, the more likely messages are to be brief, abbreviated, and possibly off-the-cuff (Werry, 1996). Grinter and Eldridge (2001) believe that teens are able to turn these limitations to their own advantage. Crystal (2008b, p. 69) believes that abbreviations, such as the ones noticed in SMS, “began as a natural, intuitive response to a technological problem.”

Crystal (2008b) discusses at length the ergonomic difficulties with inputting text on a cellphone. “The keypad was not originally designed with language in mind ... Apart from anything else, phones keep getting smaller and smaller, but fingers stay the same size” (Crystal, 2008b, p. 65). Most smartphones now have touch keyboards or mini-keypads⁹ that have improved input in recent years; however, both still provide keys that are often too small for an individual’s fingertips, and so fingers may accidentally press the wrong key. Further, mobile devices—even ones with touch or mini-keypads—often are unable to display a full keyboard. Instead one must select from several keypads: one for letters, one for numbers, one for punctuation—thus requiring users to press more keys to enter single characters than on a full computer keyboard. Baron (2008) explains that although one keystroke is required for an apostrophe on a full computer keyboard, many cellphones require users to press four keys to input the apostrophe. To add another complication, many mobile devices use predictive texting “where the phone uses a dictionary to guess which word you want to say” (Crystal, 2008b, p. 67). Sometimes the phone guesses the wrong word and users accidentally (or reflexively) accept incorrect words, sometimes because scrolling through a long list of

⁹ The term *mini-keypad* is referring to a tiny version of a full computer’s keyboard. In other words, these smartphones no longer force the user to use the number keys found on phones to type letters, where one might have to press the 2 key three times to type a C.

possible words to find the correct word may be tedious. So users may have to go back and retype the words correctly. Subjects from Thurlow's (2003) study indicated that they found predictive texting difficult to use and annoying.

The size of the screen can also influence message production and viewing ease. Mobile devices such as cellphones offer a very small viewing area, whereas full-sized computers offer more real estate within which to compose and view messages. However, even on a traditional computer, some media are relegated to a small window size due to other factors. For instance, some chat is offered through a webpage (e.g., AOL chat) or larger application (e.g., WoW chat) where the chat window itself assumes a smaller portion of the overall application window. In WoW, you may expand the chat window quite a bit; however doing so, means that you conceal the animated game play, which can severely limit player performance. Therefore, WoW chat windows are often kept rather small.

Quoting and linking

Some media, such as email and many forums, offer interlocutors the opportunity to quote or link back to another interlocutor's comment. It "incorporates and juxtaposes (portions of) two turns – an initiation and a response – within a single message" (Herring, 2001, p. 620). This feature helps users manage coherence better than what is possible in some non-quoting media, such as chat. Instead of having to self-determine which response is logically paired with which earlier comment by sorting through multiple comments, as can be the case in chat, media with quoting features will allow interlocutors to include snippets of earlier comments within their own responding messages, so that the pairs (message and response) are linked visually on the page. "These adaptive strategies compensate for a lack of

simultaneous feedback in one-way computer communication systems by providing explicit mechanisms for speaker change” (Herring, 2001, p. 620).

Media such as IM and chat do not often provide such affordances and users may be expected to manually copy/paste previous messages. Since these two media are synchronous and often subject to unceasing message scrolling, there may be no opportune moment to copy and paste prior messages without falling behind in the conversation. Furthermore, if there are also message length limits, copying another’s response into your own message may not be practical—there may be no room for a response once the original message has been copied.

Conclusion

Media that might be more likely to exhibit cyberlanguage may possess the following characteristics:

- synchronicity (as opposed to asynchronous media) because it is immediate and thus may require faster, more speech-like responses and provide less time for planning and editing messages, both which may lead to more abbreviations, disfluencies, and play,
- higher participant scale, such as N:N, because there may be a greater need to keep up with the conversation leading to more abbreviations and play,
- low message persistence in terms of both visibility and re-use because the conversation may appear more ephemeral and therefore more speech-like, which may lead to more disfluencies,
- more public/open conversations rather than private because this may encourage participants to perform and play more,

- greater anonymity which may contribute to more uninhibited behavior, performance, and play,
- shorter message length which may encourage abbreviation,
- more difficult input devices that may make message composition tedious and deter users from planning and editing messages, and instead encourage them to abbreviate more.

These observations contribute to the formation of the research questions for this dissertation study.

Cybermedia descriptions

Forums

Discussion forums—also referred to as bulletin boards, bulletin board systems, BBS—are asynchronous. So there is “time to read, understand, and respond, without the pressures of real-time interaction” (Crystal, 2006, p. 267). Hence, interlocutors may spend more time planning and revising messages before posting; as a result, messages may be more formal in tone and use more standard grammar, punctuation, and spelling. Additionally, one can copy existing quotes (either manually or through some affordance offered by the source) from other users’ postings and paste them into a new post (Herring, 2007), pairing responses with earlier comments to achieve a sense of immediacy leading to a more conversational tone.

Message length restrictions may be nonexistent, however some forum software may require the user to enter text into a text box that may not expand as one types. So entire message bodies may not be visible during composition, thereby making composition more

difficult and prone to error. However, standard keyboards can be used for editing, and viewing messages is only limited by the user's screen size and the size of the composition and viewing frame within the webpage.

Conversations typically begin as 1:N, with one interlocutor posing a question or topic for discussion. If the topic is attractive to other interlocutors and inspires them to respond, the conversation can easily turn to N:N discussion. Messages are often listed in the "temporal sequence in which they were posted, or grouped into 'threads' according to subject line" (Herring, 2002, p. 117).

Forum participants are often not previously acquainted (Herring, 2002). They may have to register for a user account in order to post or view messages; however, the level of privacy is variable because some forums may not require login to view messages, making the forum, in effect, public/open. Even though a user account and login may be required for entrée, some are not moderated and do not maintain a list of subscribers and so accountability may be reduced (Herring, 2002). Even if the forum is moderated it may be difficult for an administrator to keep up with a high volume of user accounts; use of pseudonyms instead of personally identifying userids may also make moderation difficult.

Forums tend to have a greater degree of message persistence, both in terms of visibility and re-use, than chat or IM. Messages do not scroll out of any buffers and conversations may continue to be posted on websites for many years.

Discussion forums are used for a variety of purposes including everything from more formal discussions (e.g., online class discussions) to more colloquial discussions (e.g., hobbies or personal interests). Given the variety of topics and the relative degree of

anonymity, it is possible that interlocutors may feel occasionally inspired to creatively and playfully reshape language.

Email

Email and email lists (such as listservs) are asynchronous (Baron, 2003; Crystal, 2006; Herring, 2002). Thus, there may be more time to plan and revise before sending messages, possibly making them linguistically complex (Herring, 2002). However, Crystal (2006, p. 155) speculates that “the speed and spontaneity with which e-mails can be written” may reduce the likelihood of reflection. Furthermore, many emails are personal in nature, thus invoking a more colloquial tone (Herring, 2002). Danet (2001, p. 57) says that email is characterized by “speech-like features”; people often write as they talk so email may feel “dynamic, interactive, ephemeral.”

“There is enormous variation in the language style used in email, determined by such variables as age, computer experience of user, and function” (Baron, 2003, p. 77). Email is used for a wide variety of purposes, thus messages can vary greatly in terms of formality (Crystal, 2006); overall, email tends to be less formal than other forms of edited writing (Hård af Segerstad, 2002; Herring, 2001). In Gains’ (1999) analysis of business email, the majority possessed a semi-formal tone and employed standard English; however, the university emails Gains analyzed were more phatic than the business emails. Crystal (2006) notes that, because emails can be so easily deleted, they may not seem as official as print documents, and thus composition style is often viewed as more informal. It is the “spontaneity, speed, privacy, and leisure value” of email that bestows greater informality than traditional writing styles (Crystal, 2006, p. 133).

Earlier emails were modeled on handwritten letters, and still today sometimes retain some of the structure of letters, with some form of salutation and closing (Herring, 2002); however, use of such openers and closers is highly variable (Crystal, 2006). In 1999, Gains reviewed 119 emails from both businesses and universities and found that 92% of the 62 business emails had no opening greeting at all. Forty-two percent of business emails closed with just the sender's name, 40% closed with some variant of *thank you*, and 8% had no closing at all (Gains, 1999). Out of the 54 university emails, 63% had some type of greeting, 24% closed with just the sender's name, and 9% had no closing at all (Gains, 1999). Waldvogel's 2007 analysis of 515 email messages at an educational institution and a manufacturing plant revealed that 53% of the 121 manufacturing emails opened with a name and a greeting word (e.g., *hi*), 25% opened with a name only, 17% had no greeting at all, and 5% opened with just a greeting word. Seventy-three percent of the manufacturing plant emails closed with a thanks or farewell word (e.g., *cheers*) plus a name, 15% percent closed with just a name, 10% with no closing at all, and 2% closed with a thanks only. Fifty-nine percent of the 394 educational institution emails contained no greeting at all, 21% contained a first name only, 15% contained a greeting word plus a name, and 5% contained just a greeting word. Thirty-eight percent of the educational emails closed with just a name, 34% contained no closing at all, 23% contained a thanks or farewell word plus a name, and 5% contained a thanks or farewell word only.

Email may be a 1:1 or 1:N form of communication (Baron, 2003). Email lists, where a single email address represents a group of people who may send messages to the entire group by using that singular address, permit 1:N and N:N conversation and are thus similar to forums in this regard (Baron, 2003). Messages tend to arrive in the order they were sent and

most email programs order incoming messages chronologically in the client's inbox.

However, users can edit the settings of their email programs to order messages by thread, sender, or other options, such as user-created tags. Users may also choose to modify email list settings so that they receive a digest of messages—several messages included in one email per day or week, for example (Herring, 2002).

In many cases, email messages may be kept indefinitely until the user decides to delete them, giving email high persistence (Herring, 2007). Even if a user chooses to delete a message, it may not be completely eradicated because a copy may still reside on the message server (as in the case of IMAP protocols for example). The Enron scandal during the early 2000s is an example of how deleting messages from an active view of one's inbox on one's personal computer does not necessarily delete the messages permanently. Copies of email list messages typically reside on a server until a system administrator chooses to flush them. Sometimes archives of these messages are made available for review on traditional webpages (Herring, 2002). Additionally, users can choose to archive messages in their client to a separate file so that they no longer take up mailbox space but are available later if needed.

Although many people choose email addresses that convey something about their identity, particularly in professional settings, some do not. And although most email users tend to know, in some fashion (either virtually, by phone, or in a face-to-face context) the person they are corresponding with (Wellman, 2004), this may not always be the case, particularly where spam, spoofing, and email lists are concerned. So recipients may be confronted with messages from senders they do not know, or do not think they know, and have no way of identifying. Thus if a list has many subscribers who have never met, there may be less accountability on the list than in an email exchange between two friends or

colleagues (Herring, 2002). However, email and email lists may have higher accountability than forums or chat because many email interlocutors will have some familiarity with each other—e.g., almost 92% of subjects in Cho’s (2010) email study had met face-to-face; so there may be greater potential for intimacy and immediacy than with forums. Email is less anonymous than other forms of computer-mediated communication (Herring, 2002).

“Email is, in principle, not intended for public view” (Baron, 2003, p. 77). Emails are intended for the eyes of the person or persons listed in the To: or Cc: fields. However, one may list someone in the Bcc: field unbeknownst to the person or persons listed in the To: and Cc: fields. Also, a message can be easily forwarded to others, thus sharing the message with someone that the original sender did not intend (Crystal, 2006). Messages may also be copied and pasted into new documents or they may be edited and sent to other people, all without the original sender’s knowledge. “The willing surrender of control over one’s written or spoken output is not in itself novel: journalists, for example have long been used to having their copy altered by senior editors before it appears in print” (Crystal, 2006, p. 127). “But e-mail permits the extension of such practices to a very wide range of communicative behaviours previously immune to such ‘interference’, and the consequences have yet to be explored” (Crystal, 2006, p. 127). So privacy in email is variable and unpredictable.

Email messages can be of any length. Crystal (2006, p. 119), in analyzing 50 personal emails he received, found considerable variation in length of messages, anywhere from 6.56 lines of text per message to 30.65 lines, with an average of 10.9 lines. Hård af Segerstad (2002) found an average message length of 63.71 words in the 183 messages he analyzed in 1998. Cho’s (2010) analysis of 197 messages revealed an average of 98.88 words per message. Paragraphing is used but paragraphs tend to be short (Crystal, 2006).

Composition and viewing of messages on standard computers may be relatively easy with a full computer keyboard and a reasonably large screen. However, some web clients may offer small, unexpandable text boxes for composition, and if mail is viewed and sent on a mobile device, composition and viewing may become more difficult.

Quoting is supported in most email programs. Interlocutors may even interleave specific responses within the original message, pairing responses with the prompts for those responses. Crystal (2006, p. 124) refers to this as “framing” and he says this technique is not like anything else in traditional language use.

SMS

Short message service (SMS) or text messaging is the act of sending messages with mobile devices like cell phones (Spagnolli, 2012). Messages may also be sent from web forms to cellphones, but most often messages are sent between mobile devices. Crystal (2008b) and others refer to those who send text messages as *texters*.

SMS is asynchronous and primarily 1:1 (Spagnolli, 2012), but 1:N communication is possible. Messages are listed in the order in which they are received. The portability and low cost of SMS make it very appealing (Crystal, 2008b). Whereas IM requires one to be in front of a computer to communicate, SMS does not—a texter can be on the move (Crystal, 2008b; Baron, 2013). Baron (2008) considers it a time-saving way to communicate because one can quickly pop off a short message. Crystal (2008b, p. 30) claims that “text messaging seems to have increased our expectation that we are mutually accessible,” in turn leading to expectations for quick responses and more constant communication. SMS is often used to coordinate activities (Crystal, 2008b; Grinter & Eldridge, 2001; Ling, 2005), as well as

monitor political polls, receive updates from political candidates, and vote on websites and TV shows and receive updates from them (Crystal, 2008b).

SMS affords more privacy than phone calls in two regards: (a) it permits covert communication in places where an audible conversation would be inappropriate and (b) discrete communication when interlocutors do not wish to be overhead (Crystal, 2008b; Grinter & Eldridge, 2001). It appeals to those who do not want to waste time “engaging in linguistic hand-shaking” such as greetings (Crystal, 2008b, p. 96). “Messages are typically sent between people who know each other well. This means that the language will be intimate and local, and make assumptions about prior knowledge” (Crystal, 2008b, p. 52). When a message is received, the sender’s phone number is what identifies the sender to the recipient (as opposed to a traditional userid or pseudonym as in the case of the other media discussed). If the recipient knows the sender and has entered his/her contact information, including his/her name, into the in-phone phonebook, then when the recipient receives the message, instead of the phone number appearing, the sender’s name (as entered into the phonebook by the recipient) will appear.

Messages are stored in the phone and can be retrieved easily (Spagnolli, 2010). Like emails, messages can be saved indefinitely until the user chooses to delete them; however, some phones permit a limited number of messages to be stored in memory. Messages can be forwarded and downloaded to computers (depending on the phone’s capabilities), so there is some ability to replicate and pass on messages to parties not originally intended to receive them. Thus, although text messages are largely private between sender and receiver, they can be made more public.

Many text messaging services allow only a limited number of characters to be sent, usually 160 (Crystal, 2008b, Hård af Segerstad, 2002; Thurlow & Poff, 2013). In Hård af Segerstad's (2002) analysis of 1,152 messages—some of which were personal messages and messages of family and friends—there were 14.77 words per message (64 characters per message) in the Swedish corpus and 13 words per message (78 characters per message) in the German corpus. In her analysis of 191 text messages, Baron (2008) found that messages averaged 7.7 words and that one-word transmissions were rare, which could be due to the cost of messaging. In other words, it may be wasteful to spend money on a one-word message (Baron, 2008). Holtgraves' (2011) study showed similar results to Baron's: average words per message was 8.11, and 90% of messages contained fewer than 17 words. Thurlow and Poff (2013) report 14 words per message (65 characters per message) on average. So text messages tend to be brief, and are not all that dissimilar from “scribbled notes” (Thurlow & Poff, 2013).

Given the tiny screen size and keypad of most mobile devices, most text messages are not particularly easy to create, and this seems to have led to a rather abbreviated linguistic style (Crystal, 2006; Spagnolli, 2010). For instance, depending on the phone, one may need to punch a given key several times for it to cycle through the available characters before it gets to the character the user wishes to appear in the message (Baron, 2003). For some users, an abbreviated style may also be more economical since some phone companies charge not by the message but by the character (Crystal, 2006). Because of the difficulty in creating messages, less time may be spent on planning and editing messages before sending them (Baron, 2003). “Compared with a formal letter or an academic essay, [text messages] are most likely shorter..., contain more language play, and are more chatty” (Thurlow & Poff,

2013, p. 179). Texters “write it as if saying it” and thus create an informal register for small talk and sociality (Thurlow & Poff, 2013, p. 177).

IM

IM is a synchronous medium (Tagliamonte & Denis, 2008). Messages scroll by in the order in which they were received. IM primarily supports 1:1 conversation (Tagliamonte & Denis, 2008). However, a user can carry on multiple, simultaneous conversations with others; each conversation will be separate with its own IM window. So if a user wishes to have conversations with five people, he must have five separate IM windows open. Paolillo and Zelenkauskaitė (2013) claim that IM conversations are more private than chat conversations. Nardi, Whittaker, and Bradner (2000, p. 82) claim that IM is “opportunistic, brief, context-rich and dyadic,” and as such participants perceived it as being interchangeable with face-to-face communication. They explain that IM is seen as faster than email because, like SMS, the formalities of greetings can be dispensed with.

Interlocutors using IM, like those using email and SMS, are more likely to be familiar with one another (Hård af Segerstad, 2002) than in forums or chat, so the degree of anonymity is low. There are some IM situations where that familiarity is situational, as in the case of Ask a Librarian IM or IM used for technical support. This familiarity, coupled with the 1:1 participant scale, may make IM more formal in tone, particularly when used for professional or academic reasons. However, Nardi et al. (2000)¹⁰ claim that IM is flexible and expressive, allowing for affective communication such as joking and relieving the stress of the workday. They found that IM is used for quick questions, clarifications, coordination

¹⁰ The Nardi et al. (2000) study took place at an Internet company.

and scheduling, keeping in touch with family and friends, and negotiating availability in other media (such as setting a time to talk on the phone) (Nardi et al., 2000). The 8,255 messages sent in a computer lab that Hård af Segerstad (2002) collected in the late 90s focused on social coordination, work-related and task-related topics, greetings, system testing (e.g., “did you get my message?”), asking what a person is doing or where s/he is, imitating the system (“fatal error!”), and phatic statements such as reprimands and encouragement. Users also enjoy monitoring buddy lists to keep tabs on who is on so they feel more connected, which Nardi et al. (2000, p 85) refer to as “awareness moments.” Some IM programs even provide support for phatic responses by offering graphic symbols that can be inserted into the message to indicate facial expression or emotion (image-based emoticons) such as 😊¹¹ so that users do not have to create their own, such as the equivalent, :-).

Like chat, IM messages appear in a limited buffer and may scroll out of view. Once the buffer runs out, messages may be lost forever. Some IM software does allow users to archive messages however. Because IM is synchronous and immediate and because messages can quickly scroll out of the buffer, careful planning and editing may not occur. Therefore, there may be more disfluencies and spontaneous play than in forums or email.

Most IM programs also impose message length restrictions on interlocutors. So messages tend to be brief and may use more abbreviations. Baron’s (2008, p. 57) review of 23 IM messages showed an average length of 5.4 words. She (2008) concludes that the shorter length makes IM more akin to speech than writing. IM interlocutors may also use several transmission lines to complete a turn (Isaacs, Walendowski, Whittaker, Schiano, &

¹¹ Taken from one of the stock set of graphic emoticons offered by the chat aggregating application Adium (<http://adium.im/>).

Kamm, 2002). Baron (2008, p. 49) refers to this as “chunking.” For example, in this hypothetical conversation, Joe continues his turn on four lines:

Joe: what i want to say...

Susan: yes...

Joe: is that i don't agree with you on this 1

Joe: it seems to me that white chocolate cant be considered real chocolate

Joe: because it doesnt include chocolate liquor

Users compose messages using a standard keyboard and so are not limited in their typing ability in quite the same way texters¹² are (Baron, 2013). However, IM users may be relegated to small text box sizes that, along with the message length restrictions, may curtail lengthy prose.

Whereas the N:N atmosphere of chat may pressure interlocutors into conversing, the 1:1 nature of IM may make it possible for IM users to treat it more like email or SMS where messages may be screened and responded to later if desired. Nardi, et al. (2000, p. 84) qualify this as “plausible deniability.” The 20 subjects in Nardi et al.'s (2000) study indicated that IM seemed less interruptive than popping into someone's office, which was viewed as more “in your face.” One participant said it “interrupt[s] them without interrupting them too much” (Nardi, et al., 2000, p. 83). Participants also felt that it afforded them greater control over not only determining when someone was available but also avoiding interruption by others.

¹² *Texters* are those who use SMS and send text messages.

Instead of conversations taking place at the convenience of the initiator, IM allows genuine social negotiation about whether and when to talk. The attentional contract can be negotiated on a more equal footing between initiator and recipient than with face to face or phone interaction. This may explain why IM is often used to negotiate availability for phone calls and face to face conversations. (Nardi, et al., 2000, p. 84)

Chat

Chat is synchronous, so conversation tends to be rapid and rather spontaneous, mimicking the pace of face-to-face communication (Werry, 1996). Chat media can support 1:1, 1:N, and N:N conversations. “Chat occurs in a broad range of contexts” (Paolillo & Zelenkauskaitė, 2013, p. 109). Examples of chat programs include IRC,¹³ MUDs and MOOs, MMOGs, AOL Chat, etc. “Chat is typically organized according to ‘chat rooms’ or ‘chat channels’” (Paolillo & Zelenkauskaitė, 2013, p. 111). Often times different scales (e.g., 1:1, 1:N, N:N) are handled with different chat channels. In WoW chat for example, 1:1 conversations are called *whispers* or *tells* (aka *pages* in MUDs/MOOs (Cherny, 1999)) and have their own channel. A small group N:N conversation is also handled via different channels in WoW, such as the *party* channel and the *raid* channel. Most chat programs provide some visual distinction between the different channels. In WoW, channels are color-coded and utterances are preceded with the channel name.

Chat discourse may appear chaotic, with disrupted turn-taking patterns (Herring, 2002, p. 121). Conversation is often colloquial and informal (Crystal, 2006; Werry, 1996). The “ephemerality, speed, interactivity, and freedom from the tyranny of materials” may encourage playful speech (Danet, Ruedenberg-Wright, & Rosenbaum-Tamari, 1997, An Inherently Playful Medium section, para. 2). N:N conversations in chat may resemble a “cocktail party in which everyone is talking at once – except that it is worse, because every

¹³ IRC stands for *Internet Relay Chat*—one of the early chat services offered.

guest can ‘hear’ every conversation equally, and all guests need to keep talking in order to prove to others that they are still involved in the interchange” (Crystal, 2006, p. 165). However, many gamers and chat aficionados become quite adept at focusing on the streams that interest them. “Topics decay quickly, making unstructured chat un conducive to extended focused discussion” (Herring, 2002, p. 121). Cherny (1999, pp. 178-179) describes chat as a “collaborative floor” composed of a “main floor” (i.e., the focal stream) and parallel “side floors” (i.e., other streams of potential interest).

Chat messages scroll by in real time, in the order in which they are received (Ferrara et al. 1991; Herring, 2002), making turn management difficult. The overlap between turns may give chat a chaotic appearance (Crystal, 2006; Davis & Brewer, 1997; Herring, 2002). For instance, in the time it takes Person B to compose a response to Person A’s earlier comment, Person C could have interjected an utterance of her own. “Conversation proceeds, in a mixture of sequence, simultaneity, and overlap” (Crystal, 2006, p. 158). However, Cherny (1999) believes that true overlap—that which is defined by one person talking *over* another as one might see in face-to-face communication—in chat is not possible. Although chat may appear chaotic, Collister (2008, p. 83) claims that “there is logic to it, there is order, and these things are observable.”

Danet (2001) and Herring (2002) suggest that the interruption and overlapping streams of conversation found in chat foster playfulness. As such, these modes invoke interactivity and involvement from users (Cherny, 1999; Ferrara et al. 1991; Werry, 1996).

Chat “users can scroll back to read earlier messages” but it is within a limited buffer size (Herring, 2002, p. 121), and at some point, the buffer ceases to afford further scroll back. When the buffer runs out, messages may be permanently lost. However, some chat software

does allow copying/pasting and some allow for archiving chat logs. It is also possible that what occurs in email scenarios is true of IM and chat, that system administrators keep logs stored on the server. So the degree of persistence is relative, but considerably more ephemeral than email, for example (Herring, 2007).

“The presence of multiple participants makes [chat] more public than private” (Paolillo & Zelenkauskaitė, 2013, p. 111). However, chat users frequently choose pseudonyms that may not include personally identifying information, which allows them to obscure their identities (Paolillo & Zelenkauskaitė, 2013; Silva, 2010). This may afford interlocutors a high degree of anonymity (Crystal, 2006; Hård af Segerstad, 2002; Herring, 2007). The use of pseudonyms often induces a masked ball atmosphere where conviviality and frivolity abound (Danet, 2001). “The culture of chat rooms, although varying according to purpose, is typically sociable, playful, and disinhibited” (Herring, 2002, p. 121). Danet (2001) says that chat spans five frames of engagement: real life, the IRC game, party, pretend, and performance. Crystal (2006, p. 175) explains that it is like “attending a perpetual linguistic party, where you bring your language, not a bottle.” “Language play is routine” and chat conversations have a “strongly phatic character” (Crystal, 2006, pp. 174-175). This masking of identity may help chatters “generate familiarity and intimacy”—“a type of language that abbreviates the physical and emotional distance between them” (Silva, 2010, p. 268).

Chat is often subjected to message length restrictions (Herring, 2007). This may lead to less complex messages and more abbreviations. “Contributions tend to be single sentences or sentence fragments; and word-length is reduced through the use of abbreviations and initialisms” (Crystal, 2006, p. 162). Ferrara et al. (1991, p. 12) say messages are “part

postcardese,” “part telegraphese,” and “part headlines.” In “a sample of 100 direct-speech contributions taken from published log data,” Crystal (2006, p. 162-163) discovered “an average of 4.23 words per contribution, with 80 per cent of the utterances being 5 words or less,” which Crystal concluded allows for a more “real-time dynamic.” Out of a year’s worth of chat logs (about 25MB), Cherny (1999, p. 155) noticed that messages tended to be between five and 13 words long, and their brevity contributes to the sense of co-presence and awareness of others.

The synchronicity and N:N participant scale of chat may necessitate speed: “fast feedback is required, which demands an economy of writing to guarantee the conversational dynamicity” (Silva, 2010, p. 268). These characteristics, along with limited message length, may also result in less planning and editing of messages and more “extemporaneous composing” (Davis & Brewer, 1997, p. 29). Collister (2008) and Silva (2010) believe that chat resembles speech. “Chatgroups are the nearest we are likely to get to seeing written dialogue in its spontaneous, unedited, naked state” (Crystal, 2006, p. 176). Thus, messages tend to be rather informal (Herring, 2002).

Chat programs are often used on desktop computers, so full keyboards and reasonably large screens may be used, making viewing and composition easier than with SMS on mobile devices. However, sometimes the chat window may be a part of a larger program, as in the case of WoW chat, where the chat window assumes a small portion of the screen. It may be expanded, but to do so would result in obscuring parts of the screen that show game play, thereby making game play more difficult.

Conclusion

Table 1 below shows how each medium fares with respect to the media characteristics outlined in the previous section.

Table 1: Cybermedia and their characteristics in the most typical scenarios.

	Forums	Email	Email Lists	SMS	IM	Chat
Synchronous	no	no	no	no	yes	yes
Participant Scale (most common)	1:N, N:N	1:1, 1:N	1:N	1:1, 1:N	1:1	1:1, 1:N, N:N
Message Permanence: Visibility	extended viewing	extended viewing	extended viewing	extended viewing	limited viewing	limited viewing
Message Permanence: Re-Use	extended storage possible	extended storage possible	extended storage possible	extended storage possible	storage may not be possible or may require the user to manually set storage up	storage may not be possible or may require the user to manually set storage up
Privacy	usually open, but some forums may be closed to registered users	expectation of privacy, but not a guarantee	usually closed to subscribers, but archives may be publicly available	expectation of privacy, but not a guarantee	expectation of privacy, but not a guarantee	usually open, but some chat rooms may be closed to registered users
Anonymity	pseudonyms possible, but sometimes personally identifying email addresses are used	email addresses may be personally identifying	email addresses may be personally identifying	userid's are typically phone numbers, which are personally identifying to some degree	pseudonyms possible, but sometimes personally identifying email addresses are used	pseudonyms possible, but sometimes personally identifying userid's are selected
Message Length	no restrictions typically	no restrictions typically	no restrictions typically	restrictions apply, usually 160 characters	restrictions apply, dependent on the particular IM program	restrictions apply, dependent on the particular chat program

	Forums	Email	Email Lists	SMS	IM	Chat
Compositional Ease	easy: standard keyboard, text box may be small	easy: standard keyboard, expandable window	easy: standard keyboard, expandable window	difficult: mini keypad, very small text box	possibly difficult: standard keyboard, text box may be small	possibly difficult: standard keyboard, text box may be small
Viewing Ease	easy: full computer screen	easy: full computer screen	easy: full computer screen	difficult: small screen	easy: full computer screen	easy: full computer screen; difficult: if within a game
Quoting	supported	supported	supported	typically unsupported	typically unsupported	typically unsupported

Theories about Communication in Lean Media

As online, conversational media became more popular, researchers began to theorize—both by applying pre-existing theories and developing new theories—about the effects communication media might have on the level of immediacy and intimacy in conversations, and about which kinds of communication functions were better suited to online communication. Several theories emerge—or re-emerge—around these concerns: social presence (Short, Williams, & Christie, 1976), media richness (Daft & Lengel, 1986), lack of social context cues (Sproull & Kiesler, 1986), channel expansion (Carlson & Zmud, 1999), and social information processing (Walther, 1992; Walther, Loh, & Granka, 2005).

Social presence theory

Social presence theory predates most online conversational media, but it has been applied to online communication with regard to these newer media. It suggests that different constraints and affordances of media—inclusion or exclusion of a visual channel, for

example—will affect the level of intimacy possible and the degree to which an interlocutor may inject a sense of his/her own personal presence into the conversation. However, because it is difficult to determine which types of cues—e.g., eye gaze, hand gestures—prompt greater intimacy, it is difficult to determine which “communication outcomes will be affected” by the inclusion or exclusion of certain media features (Short et al. 1976, p. 59). “One can only point to the type of tasks most likely to be affected and the effects to be expected” (Short et al., 1976, p. 59). Thus, certain media may be better for certain tasks, conversational topics, or communicative purposes. For example, media that do not support the communication of face-to-face cues might be better for more task-oriented, less personal interactions. Short et al. (1976) are careful to explain that even though one might be tempted to say all negotiations are better handled face-to-face rather than on the phone, for example, communication is a bit more variable. In any given conversation, objectives may change and new ones may evolve, which will result in changes in the social context. So it is difficult to successfully classify conversations and their purposes as being better suited to certain media. Furthermore, interlocutors will have their own impressions about the media they use, including impressions about its aesthetic appeal, which will affect their perception of the level of social presence possible. No matter what degree of intimacy may or may not be possible, Short et al. (1976) explain that interlocutors can compensate for lower levels of presence and intimacy by altering their speech to make the conversation feel more “immediate”; for example, interlocutors use the pronoun *we* instead of *I* or *you* for a more inclusive feeling.

Media richness

Media richness, similar to social presence in some respects, posits that the constraints and affordances of the medium affect the level of interpersonal involvement that interlocutors can experience (Daft & Lengel, 1986). Specifically, the more opportunity for face-to-face cues the richer the medium should be, and the more likely interlocutors will understand one another (Daft & Lengel, 1986). Through tone of voice, emphasis, facial expressions, gesture, body language, and the like, it is thought that interlocutors are better able to help disambiguate messages and reach greater understanding. Thus media that incorporate spoken and visual channels are thought to be better for communicating about “equivocal” topics (Daft & Lengel, 1986), and media of low richness is presumed to be better suited for unequivocal messages and “standard data” (Daft & Lengel, 1986, p. 560).

“Face-to-face is the richest medium because it provides immediate feedback so that interpretation can be checked” (Daft & Lengel, 1986, p. 560). This approach seems to suggest that, in face-to-face conversations that would allow for the range of all possible face-to-face cues, interlocutors have few, if any, difficulties disambiguating, and that there are few, if any, opportunities for verifying interpretation in non-face-to-face situations. This may not, however, be the reality of the situation. Facial expressions, gestures, tone of voice, body language are often misinterpreted, and verifying an interpretation is a matter of interlocutor choice and initiative-taking. Liwei (2001, p. 18) explains that context can help with interpretation, and so “new usages” should not “cause confusion or miscomprehension...even though people need to get used” to these new language uses; such unorthodox constructions help “save both time and energy” and “help make electronic communication more effective.”

Daft and Lengel (1986, p. 560) believe that rich media—that which allows a broader range of face-to-face cues—are more personal while media of low richness is impersonal. Countering this idea, Reid (1991, Discourse and Moral Judgment section, para. 8) claims that “the idea that as the communication bandwidth narrows interaction should become increasingly impersonal does not hold true for IRC.” She cautions against labeling online communication as “shallow or ephemeral” and explains how chat’s invitation to experiment with identity and to act less inhibited encourages self-disclosure which can lead to greater intimacy (Reid, 1991, Reduced Self-Regulation section, para. 7).

Lack of social context clues

Sproull and Kiesler (1986, p. 1495) have concluded that “when social context cues are strong, behavior tends to be relatively other-focused, differentiated, and controlled;” but “when social context cues are weak, people’s feelings of anonymity tend to produce relatively self-centered and unregulated behavior.” In this model, richness is fixed and seems to disregard the interlocutor’s part in the conversation; rather, interlocutor behavior almost appears to be directed by the medium. For instance, Sproull and Kiesler (1986) speculate that because email is drafted in private (without the physical presence of the recipient), it encourages the sender to focus on him/herself rather than on the other person. They believe that this self-centered behavior is evident when interlocutors omit greetings but include closings (Sproull & Kiesler, 1986). Thus, with this approach, interlocutors possess no, or very little, free will to create intimacy, immediacy, and personal involvement if they are not in close proximity and visible to one another.

Channel expansion theory

Channel expansion theory suggests a medium's richness is not solely a property of the medium, nor is it fixed (Carlson & Zmud, 1999). It is also a matter of interlocutor perception. "Over time, as individuals add to knowledge bases relevant to the effective and efficient use of a channel, they will come to view that channel as increasingly rich" but these perceptions will eventually stabilize (Carlson & Zmud, 1999, p. 157). For Carlson and Zmud (1999), an interlocutor's experience with the medium as well as experience with other conversation partners shape his/her perceptions of how rich the medium is. "Individual beliefs concerning the appropriate use of a channel as well as perceptions of a channel's richness (perceived media richness) are, in part, socially constructed and therefore subject to social influence" (Carlson & Zmud, 1999, p. 156).

Social information processing

Social Information Processing (SIP) theory offers a more favorable view of online communication than these other models.

SIP rejects the view that CMC¹⁴ is inherently impersonal and that because nonverbal cues are not available in CMC that relational information is therefore inaccessible to CMC users. Rather, SIP posits that users employ the verbal characteristics of CMC to convey the relational information that would normally have been expressed through nonverbal cues. (Walther, Loh, & Granka, 2005, p. 40)

Walther (1992) explains that all people are driven by needs for affiliation, social acceptance and reward—i.e., we are social animals who desire social relationships with others. "CMC

¹⁴ CMC stands for *computer-mediated communication*.

users, just as communicators in any context, should desire to transact personal, rewarding, complex relationships and that they will communicate to do so” (Walther, 1992, p. 68). To develop these relationships, people seek out others, and online, initially develop simple impressions through the textual information communicated in cybermedia. They then test these impressions and assumptions over time by gathering more information about others. During this process, they will compensate for the missing face-to-face cues by creating textual surrogates such as emoticons. Liwei (2001, p. 18) believes this to be true of email users; they will compensate for missing cues by “employ[ing] linguistic and non-linguistic usages...such as abbreviations and emoticons.” Walther (1992) explains that these new “textual cues will become [interlocutors’] stock in trade” (p. 75) and that these compensation strategies “are more robust than can be impeded for long by computer mediation” (p. 80). Herring (2002, p.140) underscores these points:

Social meanings appear to be conveyed effectively through CMC. Users achieve this in part through creative uses of language, such as novel spellings, repeated punctuation, and ASCII¹⁵ graphics designed to convey attitude, non-speech sounds and facial expressions.

Walther’s (1996, p. 17) “hyperpersonal” model extends this idea by suggesting that sometimes online communication can surpass “the level of affection and emotion of parallel FtF¹⁶ interactions” and possibly become “more socially desirable than we tend to experience in parallel FtF interaction.”

¹⁵ *ASCII* stands for *American Standard Code for Information Interchange*. It is a character-encoding scheme and in this paper refers to the use of keyboard characters.

¹⁶ *FtF* or *FTF* is an acronym for *face-to-face*.

Conclusion

Theories that postulate low richness based on the leanness of the medium tend to “take face-to-face as the ideal medium for any communication needs, and root the richness of a certain medium on its technical properties” (Spagnolli, 2010, p. 3). These approaches fail to consider factors beyond that of the medium itself, such as other facets of the context, the purpose for communicating, the evolving nature of conversations, and the human desire for social relations. This may result in a narrow view of the types of engagement possible in cybermedia. “It is clear that simple notions of involvement or engagement based on the physicality of face-to-face dialogue (such as non-verbal signals, prosodics and paralinguistic effects) that have been the mainstay of linguistic accounts of dyadic exchanges do not take us very far in this new communication context” (Carter, 2004, p. 193). For the purposes of this study, theories that factor in these other aspects of the communication situation—primarily Social Information Processing—will be used as the backdrop for understanding the personal and social elements of online conversation.

Other Situational Variables: Genre

Medium alone cannot determine the type of language used (Spagnolli, 2010). Herring (2007) explains that medium and situation jointly influence communication. So in addition to considering the medium’s influence on language production, this study will examine other characteristics of the communication context and their potential influence on language production. These characteristics combine in particular ways and the result is a particular type of text, or genre. “Genre is a typifying concept: Instances of utterances resemble one

another and can be classified or recognised thereby” (Giltrow, 2013, p. 717). It is a “phenomenon at the interface of language and sociality” (Giltrow, 2013, p. 717).

Disciplines view genre differently, so there are many definitions of it, making it “a fuzzy concept” (Swales, 1990, p. 33). In this section, a description of each of the characteristics researchers believe comprise genre will be provided, followed by a summary of these characteristics for the purpose of arriving at a core definition that guides this research study. Characteristics that could shape the cybermedia conversations include the purpose for communicating, attributes of the communicators, communication norms and expectations, the situation at hand, the content of communication, and structure.

Purpose

Purpose is the interlocutor’s intention or objective driving his/her efforts to express him/herself. “Genres are communicative vehicles for the achievement of goals” (Swales, 1990, p. 45). These goals could include the desire to persuade, to entertain, to console, to prove a point, or, more functionally, to apply for a job, to report study findings, etc. With any communication situation, interlocutors may have multiple purposes for communicating or purposes may alternate (Beghtol, 2001; Swales, 1990). For example, within a blog post, one might see evidence of several purposes: to state an opinion, to invite a discussion of that opinion, to clarify one’s point of view, to refute conflicting ideas, etc. Shifting purposes may make it difficult to apply a singular genre label to a text (Swales, 1990).

Interlocutors

Interlocutors influence the shape of the communication; so a genre is “partially defined by its user group” (Rosso, 2008, p. 1054). The users may have a particular way of associating with one another—a particular culture—that affects text and language production. For example, faculty can be seen as participating in an intellectual culture which may result in more formal speech, more complex sentences, more structured documents when communicating their work to the intellectual community or when talking to students, so as to convey a sense of seriousness to the exchange. Gamers, however, belong to a culture of play, and so communication may be informal, less linguistically complex, and unstructured in comparison.

Norms, expectations, and the situation

In addition to culture, interlocutors within a particular communication situation may impose, in a formalized way or not, certain norms and expectations for communication. “Genre conventions signal a discourse community's norms, epistemology, ideology, and social ontology” (Berkenkotter & Huckin, 1995, p. 4). Choice of genre may depend upon these norms and expectations, as well as the particulars of the situation at hand (Rosso, 2008). For example, it may be expected that one will submit a cover letter with a résumé or curriculum vitae (CV) when applying for a job; however it is possible some hiring managers in certain fields only wish to see a résumé or CV. Some discourse communities may frown on typographical errors, misspellings, or seeming grammatical errors in all communication, including more informal communication such as email. Others may be less distraught by

disfluencies and more focused on other aspects of the communication, such as immediacy and intimacy.

Content

Content—the topic, subject matter, or meaning—also has an influence on the resulting text. It is manifested in words, phrases, and sentences (Toms, 2001). For example, the content of most cover letters includes some indication of the position being applied for, the qualifications of the candidate, and often an indication of a desire to continue the conversation about the position and the candidate’s qualifications at the hiring manager’s convenience. A recipe will contain content about cooking implements, ingredients, and actions required to prepare the dish. Knowing what genre a particular text is can provide clues as to what type of content will be found there (Rosso, 2008). For example, “knowing that a document is a recipe tells one that the document is about food preparation, even if the words food and preparation are not used in the recipe” (Rosso, 2008, p. 1053).

Form / structure

The particular form that the text takes—its structure—is also considered to be an important part of the definition of genre (Ferguson, 1994). *Form*, according to Toms (2001, para. 2) is “the visual appearance of the document such as formatting and layout.” Form and structure includes paragraphing, bullets, font sizes and bolding for headers, hanging indents for bibliographies, italics for emphasized words, blockquoting, etc. In her definition of form, Toms (2001) also includes the physical form of the text—such as whether it is a book or

pamphlet. Cover letters, for example, are often divided into various sections: the sender's address, the recipient's address, a salutation, the body of the letter, and the closing. A CV will have bulleted lists and headers. The shape of the text will trigger an interlocutor's mental model of that particular genre leading to the development of certain expectations about the text without having read the content first (Toms, 2001).

Conclusion

These genre facets or factors may also have an effect on the language used in the text. For example, one might expect to see more formal language in a cover letter and specific terms might include *sir*, *madam*, *qualifications*, *candidacy*, *position*, etc. In a research article, one might expect to see formal language, complex sentences, passive voice; vocabulary might include concepts about significance and significance testing including p values. In a recipe, one might find more active voice, fewer pronouns (if any), and terminology related to food items and cooking implements. In a chat log from a multiplayer game, one might see more informal language, more phrases than complete sentences, typographical errors and other disfluencies, and terminology specific to the game, such as place names for cities on the game map, weapon names, or spell names.

All of these facets—purpose, audience, expectations, situation, content, and form—converge in unique ways to form genres. “Genres are inherently dynamic rhetorical structures...and...genre knowledge is therefore best conceptualized as a form of situated cognition” (Berkenkotter & Huckin, 1995, p. 3). Genre may be summarized as a text type that:

- has a specific communicative purpose(s),

- is intended for use by a specific audience or community,
- is governed by expectations and customs of use,
- is situated in a particular context,
- focuses on a particular topic(s),
- and possesses a specific set of structural characteristics (or form).

As such texts of a certain genre may demonstrate a characteristic vocabulary and linguistic style.

Cyberlanguage and Its Characteristics

In online communication situations, “interlocutors refashion general English into abbreviated and sometimes pictographic representations of existing concepts where layers of meaning are packed into a few keystrokes” (Christopherson, 2013, Online Communication section, para 1). Cyberlanguage is a “medium of writing” that “present[s] itself as speech” (Nunes, 1997, p. 168). It is unique and “must accordingly be seen as new species of communication” (Crystal, 2006, p. 51). It is a “fourth medium,” neither speech, nor writing, nor signing (Crystal, 2006, p. 272). It may be conceived of as a “hybrid register”¹⁷ (Ferrara, et al., 1991, p. 10) or an “amalgam” of both speech and writing (Baron, 2003, p. 98), exhibiting characteristics of both. Cyberlanguage is interactive and features “heavy involvement...traditionally associated with oral language and face-to-face interaction”

¹⁷ *Register* is a language “variety according to use” (Halliday, 2007b, p. 7). It is the choice of words, phrases, and grammar for the purpose of communicating within a particular situation and for a particular genre. Halliday (2007b) further defines register as being affected by the *field*, *mode*, and *tenor* of discourse. *Field* refers to subject matter and situation, and determines content (Halliday, 2007a). *Mode* is the channel of language activity, usually some type of speech or writing (Halliday, 2007b). It “influences the speaker's selection of mood (what kind of statements he makes, such as forceful, hesitant, gnomic, qualified or reassertive; whether he asks questions and so on)” (Halliday, 2007a, p. 113). *Tenor* refers to tone, style, and formality that derive from the relationships between the interlocutors.

(Ferrara et al., 1991, p. 22). The immediacy of the online communication encourages interlocutors to mimic speech in certain ways (Crystal, 2006). So interlocutors write as they talk (Thurlow, 2003). Ling (2005, p. 347) qualifies SMS as “an extension of verbal interaction,” akin to speech. Yet cyberlanguage—like other forms of writing—allows for elaboration and expansion (Ferrara et al., 1991). Cybermedia provide, even if only for a few moments (as with chat or IM), some time for reflection that “allows users to express more precisely what they mean” (Herring, 2002, p. 140). Although less expressive than face-to-face communication, cyberlanguage is more expressive than traditional writing (Herring, 2002). It has been shown to be more lexically dense than speech but less dense than traditional writing (Yates, 1996).

Some differences in speech and writing are listed in Appendix B: Differences between Speech and Writing. However, speech and writing “do not form a simple dichotomy; there are all sorts of writing and all sorts of speech, many of which display features characteristic of the other medium” (Halliday, 1985). Halliday (1985) explains that modern technology is blurring the distinction between speech and writing. “Depending on the technology used, different forms of online communication are located at different points along a continuum from situations which elicit or facilitate the most writing-like use of language at one end, to those which elicit or facilitate the most speech-like use at the other end” (Danet, 2001, p. 16). For example, communication in asynchronous media, such as email, may fall more on the writing end of the spectrum because there may be more time for composing and editing messages; and communication in synchronous media, such as chat, may fall more on the speech end of the spectrum because these media are more immediate like face-to-face conversation (Danet, 2001). In any case, although synchronous media are

more immediate than asynchronous, they will not achieve the immediacy of face-to-face or even phone conversations because, as with all cybermedia, there is little opportunity to include authentic (i.e., not surrogate) face-to-face cues (Crystal, 2006). In face-to-face conversations, prosodic and proxemic cues are often reflexive or involuntary. In online communication, they have to be purposively “coded on to the text if they are included at all” (Ling, 2005, p. 347). This means that there is a certain artificiality to the language, keeping it apart from speech.

Crystal (2006, p. 258) suggests that online language may be emerging as a “distinctive variety...with characteristics closely related to the properties of its technological context as well as to the intentions, activities, and (to some extent) the personalities of the users.” Language depends on situation and context (Ferrara et al., 1991; Hård af Segerstad, 2002; Rúa, 2007; Shortis, 2007).

The need to keep up with the unceasing flow of conversation and the desire not to keep other interlocutors hanging in chat, for example, has been suggested as motivation for brevity (Werry, 1996). In describing SMS, Thurlow (2003) explains that there is a need for brevity and speed in communication which results in abbreviations. “In response to constraints on time, memory, and general effort, those who engage in IWD¹⁸ often use syntactically reduced forms, abbreviations, shorthand symbols, and terse phrasing, possibly modeling such features of IWD on those of other registers with severe constraints on space, such as telegraphese, headlines, or the postcard register” (Ferrara et al., 1991). Media characteristics such as synchronicity, message length restrictions, and opportunities for many-to-many conversation may spur interlocutors to abbreviate, for example.

¹⁸ *IWD* stands for *interactive written discourse*.

In cybermedia, interlocutors are not usually physically proximate with one another, so face-to-face cues such as gesture, facial expression, tone of voice are absent. This inability to see and hear face-to-face signals may prompt users to create textual surrogates for them. According to Werry (1996, p. 58), there is “an almost manic tendency to produce auditory and visual effects in writing, a straining to make written words simulate speech.” “Users compensate textually for missing auditory and gestural cues” making the language “richly expressive” (Herring, 2001, p. 614). To compensate, users code face-to-face information onto the text (Ling, 2005). In other words, “spelling is creatively manipulated in order to reproduce particular sounds” and “punctuation, in particular, is used to act as a channel for the expression of feelings” (Carter, 2004, p. 193). “Capitalisation, asterisks and exclamation marks are exploited to underline what both participants frame as a type of interaction which cannot pass without an overt expression of emotion or uses of voicing” (Carter, 2004, p. 193). “The language produced...demands to be read with the simultaneous involvement of the ear and eye” (Werry, 1996, p. 58).

Thurlow (2003) explains that the desire to redress the lack of face-to-face cues may override the need for brevity in some instances. “Linguistic economy” may also be sacrificed in favor of the need to attend to social aspects of the conversation, and so surrogate face-to-face cues may be used to introduce phatic communication into conversations (Cho, 2010). Surrogates may also help clarify message meaning (Varnhagen et al., 2010) or convey illocutionary force (Dresner & Herring, 2010). Thus, “language is transformed due to the need for economy, on the one hand, and the need to be expressive and convey one’s feelings, on the other” (Silva, 2010, p. 267).

To satisfy the goals of creating a sense of immediacy, speedy communication, and reparation for missing face-to-face cues, interlocutors may manipulate “typography (keyboard symbols, e.g., using @ for the letter *a*), orthography (alphabet and spelling, e.g., changing the spelling of *please* to *plz*), morphology (word-formation, e.g., adding the suffix *-ers* to *lol* to create *lolers*, those who laugh out loud), and syntax (combining words into utterances/sentences, e.g., omitting parts-of-speech)” (Christopherson, 2013, Online Communication section, para 1).¹⁹ What follows is a catalogue of cyberlanguage features, drawn from a review of prior research of online language.

Abbreviations

The preference for brevity has been presumed to lead to a variety of abbreviations (Ferrara et al., 1991). Abbreviations fall into two categories. They either eliminate letters or punctuation from a word or reduce the number of keystrokes.

Acronyms are an example of the former; they eliminate letters from a phrase except the initial letters of each word. Words written in lowercase that typically appear in upper case in general English texts, such as the pronoun *I*, are examples of the latter. Lowercasing requires only one keystroke rather than the two that would be required—the Shift key plus the letter key—for capitalization. (Christopherson, 2013, Online Communication section, para. 5)

The latter represents “economy of effort” (Herring, 2001, p. 617). Abbreviations found in cyberlanguage include acronyms, shortenings, clippings, single-letter forms, letter homophones, number homophones, symbolic substitution, punctuation omission, and non-

¹⁹ The definitions for typography, orthography, morphology, and syntax were taken from Herring (2012).

standard use of lowercase. In cyberlanguage, abbreviations reflect “how language adjusts to the particular constraints of these new media where speed and conciseness are of prime importance” (Rúa, 2007, p. 157).

Acronyms / initialisms

An acronym is a type of abbreviation that reduces a phrase to the initial letters of the words it contains and/or initial letters of the syllables within the words it contains. It is typically pronounced as a word (Crystal, 2008b; Quirk, Greenbaum, Leech, & Svartvik, 1985). Initialisms (sometimes referred to as alphabetisms) are essentially the same but are not pronounced as a word (Crystal, 2008b; Quirk et al., 1985). For example: *NATO* for *North Atlantic Treaty Organization* is an acronym often pronounced as *naytoh*; while *RSVP* for *répondez, s'il vous plait* is an initialism where each letter is spelled out. For convenience, the term acronym will be used primarily to indicate both acronyms and initialisms. Examples:

<i>lol</i>	laughing out loud
<i>wth</i>	what the hell

Shortenings

Shortenings are abbreviations where syllables or parts of a word are removed from the beginning, middle, or end of a word (Crystal, 2008b; Rúa, 2007). For the purposes of this study, shortenings also include vowel and consonant reduction. Vowel and consonant reduction are types of abbreviations where vowels or consonants are omitted. With consonant reduction—which appears less frequently than vowel reduction—often double-medial

consonants are removed (Crystal, 2006). This is because “consonants carry much more information than vowels” and the removal of too many consonants may result in unintelligible words and phrases (Crystal, 2008b, p. 26). Crystal (2008b) claims that interlocutors wish to be understood, so they reduce when it makes sense. For example, compare this sentence: *ths sntnc hsnt gt ny vwls* (a vowel-free sentence) with its consonant-free corollary: *i eee a o a ooa* (examples taken from Crystal, 2008b, p. 217). Both sentences mean “*this sentence hasn’t got any vowels*” but the vowel free version is more decipherable. Examples of shortenings include:

<i>gd</i>	good
<i>pls</i>	please
<i>msg</i>	message
<i>imedtly</i>	immediately
<i>puter</i>	computer
<i>prolly</i>	probably
<i>app</i>	application

Clippings

Clippings are a type of abbreviation where the final letter is dropped from the word (Crystal, 2008b). These are considered distinct from shortenings. Examples:

<i>comin</i>	coming
<i>goin</i>	going

Single-letter forms

A single-letter form is a type of abbreviation where only a single letter is used to represent an entire word. Parts of acronyms should not be confused with single-letter forms. For the purposes of this paper, acronyms and single-letter forms are distinct; to be an acronym, the term must contain more than one letter and represent a phrase not a single word. A single-letter form may also be a letter homophone (see definition below), but not all single-letter forms are letter homophones. For example, *H* for *heroic* is only a single-letter form—the sound of the letter *H* does not match the pronunciation of *heroic*—but the letter *c* for *see* is both a single-letter form and a letter homophone because the sound of the letter *c* does match the pronunciation of the word *see*. Examples:

<i>D</i>	defense
<i>H</i>	heroic
<i>b</i>	be
<i>c</i>	see

Letter homophones

Letter homophones are another type of abbreviation where a letter is substituted for its sound (Silva, 2010). They may stand in for the sound of an entire word or a sound within a word (e.g., a syllable). In effect, the pronunciation of the letter matches the pronunciation of the word or word part. When a letter homophone is used to signify the pronunciation of an entire word, it is also a single-letter form (e.g., *r* for *are*). Examples:

<i>b</i>	be, bee
<i>c</i>	see
<i>r</i>	are
<i>bhold</i>	behold

Number homophones

Number homophones, like letter homophones, substitute a number for the sound of an entire word or a sound within a word (e.g., a syllable). The pronunciation of the number matches the pronunciation of the word or word part. Examples:

8	ate
<i>gr8</i>	great
2	to, too

Symbolic substitution

Symbolic substitutions are abbreviations where a non-alphabetical symbol is used to signify a word or concept in a non-traditional or uncommon way. Examples:

???	to signify confusion
<i>apples > bananas</i>	apples are better than bananas
2	to, too

Conjunctions and disjunctions

A “conjunction is the process by which two concepts are combined as equals in a new concept” (Sager, 1990, p. 66). A slash is substituted for “and” to signify an AND condition (Christopherson, 2013).

A “disjunction is the process by which the extensions of two or more concepts are combined into a new superordinated concept. It presents two alternatives as a single concept and is therefore an either/or relationship” (Sager, 1990, p. 67). A slash is substituted for an “or” to signify an OR condition (Christopherson, 2013). Examples:

<i>will pull back to here / fight on stairs</i>	conjunction (and)
<i>morning / afternoon all</i>	disjunction (or)

Punctuation omission

When typically expected punctuation—such as an apostrophe in *don’t*—is omitted—as in *dont*—this is punctuation omission. Creating punctuation, such as apostrophes, on mobile device keypads may be difficult (Baron, 2008). Often an interlocutor has to cycle through a few keys or keypads (as in the case of some touch screens) to type the punctuation mark. Examples of punctuation omission:

<i>dont</i>	don’t
<i>ive</i>	I’ve

Non-standard use of lowercase

Many researchers (such as Baron, 2008; Danet, 2001; Ferrara et al., 1991; Werry, 1996; Yongyan, 2000) have identified a preponderance of lowercase in online communication. The initial letters of many utterances, proper names, and the pronoun *I* may not be capitalized. Examples of lowercase include:

<i>i dunno</i>	I don't know
<i>going to miami this wkd</i>	going to Miami this weekend

Surrogate prosodic cues

Prosody is one type of face-to-face cue; it reflects the aural qualities of face-to-face speech and includes “vocal variations in pitch (intonation), loudness (stress), speed, rhythm, pause, and tone of voice” (Crystal, 2006). It can also include emphasis of certain words or phrases (Crystal, 2006).

Surrogates found in cyberlanguage may be thought of as “spoken-like spelling” (Hård af Segerstad, 2002, p. 215) or “eye dialect” (Shortis, 2007, p. 5). This “innovative spelling” may be “closer to the pronunciation than the traditional spelling is” (Hård af Segerstad, 2002, p. 149). For example, prosody may be expressed with “exaggerated use of spelling and punctuation, and the use of capitals, spacing, and special symbols for emphasis” (Crystal, 2006, p. 37). “The use of exclamation marks, repetition of letters or punctuation for emphasis, and use of capitalization for emphasis imitate the effects of prosody in spoken language” (Cho, 2010, Comparison of Memos and Email section, para. 4).

While abbreviations may be viewed as a way to economize effort by saving keystrokes, many surrogates may be viewed as the antithesis of effort-economizing word

creation techniques (Hård af Segerstad, 2002). For example, letter duplication—which may signify the elongation of sounds—is a process where additional characters are added to a word. Cho believes the inclusion of prosodic surrogates demonstrates a certain “expressivity.” Surrogate prosodic cues may include onomatopoeic expression, phonetic respellings, offsetting punctuation, all caps, letter duplication, and punctuation duplication.

Onomatopoeic expression

For the purpose of this study, onomatopoeic expression includes sound effects, human vocalizations, and other noise. Silva (2010, p. 269) explains that onomatopoeic expression can show “connivance, irony, complicity, solidarity and the need to read the utterance in the non-literal sense.” Examples include:

arrgghhh!

splat!!

muahahahaha!

hmmmm

grrr

hahaha

ding

pfft

Phonetic respellings

Phonetic respellings are “shortened homophones of genuine words” (Rúa, 2007, p. 142). Words are respelled to emphasize one or more sounds within the word, often to signify colloquial or regional dialect. For example, *tunez* for *tunes* uses a *z* instead of an *s* to draw attention to the *z* sound at the end of the word when it is spoken. Phonetic respellings include elisions where two or more words are combined to convey their elided pronunciation, such as

gonna for *going to*. Varnhagen et al. (2010) suggest that elisions also speed up typing, making them a form of abbreviation also. Other examples include:

<i>wot</i>	what
<i>nuff</i>	enough
<i>dewd</i>	dude
<i>gotta</i>	got to
<i>whatcha</i>	what are you

Offsetting punctuation

Punctuation placed on either side of a word or phrase may help to emphasize that word or phrase (Crystal, 2006; Dürscheid & Frehner, 2013). Asterisks, underscores, brackets may all be used in this capacity. Examples:

<i>*that* is not a good idea</i>	<<JOE>>
<u><i>that</i></u> <i>is not a good idea</i>	{{{JOE}}}

All caps

Words typed with all capital letters, that wouldn't ordinarily be capitalized, may represent attempts at emphasis or possibly yelling (Cherny, 1999; Crystal, 2006; Silva, 2010). For example, capitalizing *that* in *THAT is not a good idea* may be for the purpose of emphasizing how much *that* is not a good idea.

Letter duplication

Interlocutors may duplicate letters in a word, possibly to mimic the elongation of sounds to convey emphasis (Crystal, 2006; Dürscheid & Frehner, 2013; Silva 2010).

Examples include *arrgghhh!* (also onomatopoeic expression) and *NOOOOOO!!!!* (also punctuation duplication).

Punctuation duplication

As with letter duplication, interlocutors may choose to duplicate punctuation to convey urgency, excitement (or extreme emotion), and emphasis (Crystal, 2006). An example: *NOOOOOO!!!!* (which also includes letter duplication).

Combinations of punctuation marks may signify surprise and confusion simultaneously, such as *what?!?!?* (Crystal, 2006). Hyphens and periods may be duplicated to signify changes in tempo, pauses, dramatic effect, or speech trailing off (Baron, 2008; Crystal, 2006; Yongyan, 2000). Examples: *it isn't going well....* or *be careful---that one is a doozy.*

Surrogate proxemic cues

Proxemics are face-to-face cues that include facial expression, gestures, and body language (Crystal, 2006). Because most cybermedia—particularly those described earlier in this paper—do not permit the demonstration of these cues, interlocutors may create surrogates for them. As with surrogate prosodic cues, surrogate proxemic cues may require more typing, making them also less directed toward economizing effort than abbreviations.

One of the most well known surrogate proxemic cues is the emoticon, used as a substitute for facial expression (Crystal, 2006). Others include emotes, symbolism for pointing, and pictograms.

Emoticons

Emoticons are “combinations of keyboard characters designed to show an emotional facial expression” (Crystal, 2006, p. 39). Dresner and Herring (2010) argue that emoticons indicate illocutionary force and are thus pragmatic. Lo (2008, p. 597) refers to them as “quasi-nonverbal cues” because they help interlocutors disambiguate the emotional and attitudinal content of messages. Emoticons are generative and may help interlocutors inject a sense of self into the conversation (Walther & D’Addario, 2001). Examples:

:~)	smile
:-(frown
^^	raised eyebrows
O.O	wide eyes

Emotes

Emotes are narrative pieces of text that attempt to convey the speaker’s behavior (i.e., “virtual action”) or state of being (Herring, 2012). Some emotes function as surrogates for body language, such as nodding in agreement (Cherny, 1999). Some cybermedia, such as MUDs, MOOs, and MMOGs, provide stock emotes. For example, in WoW chat, stock emote commands may be included in conversation by typing a slash followed by the verb for the

action, such as */cry*. If a gamer types */cry*, his/her userid plus the word *cries* will be printed to the screen (e.g., *Laura cries*). When there are no stock emote commands for a particular behavior, interlocutors may create their own, and may use the stock emote command punctuation (e.g., like the slash in WoW) to mark the text as an emote. Offsetting punctuation marks may be used to mark an emote (such as asterisks) so they resemble stage directions (Werry, 1996). Examples:

<i><John grins></i>	<i>nods</i>
<i><runs away in terror></i>	<i>/smiles</i>

Pointing

Using punctuation to create an arrow used to indicate pointing to one’s self or others in the conversation is another surrogate proxemic cue. The arrow is positioned so it points to a userid. Pointing is a special form of emote. Examples:

<i>Bob <-- is dandy</i>	Bob is dandy
<i>Superdude <== not interested</i>	Superdude is not interested

Pictograms

“When visual shapes, or pictures, are used to represent objects or concepts, they are known as pictograms” (Crystal, 2008b, p. 38). They are a type of “computer art” (Crystal, 2008b, p. 39). Letters, punctuation, and numbers are used to create pictures. Examples:

<i>@}-‘-,‘---</i>	a rose (Werry, 1996)
-------------------	----------------------

<\/>
(0v0)
^v^
()
W-W

an owl (personal email received by the
researcher)

Other features

Cyberlanguage includes other features that do not fit neatly into the abbreviations, surrogate prosodic cues, or surrogate proxemic cues categories. These include misspellings and typos, repairs of disfluencies, addressivity, reduplication, and word-formation/word-creation processes.

Misspellings and typos

A misspelling can be defined as a word incorrectly spelled because the interlocutor did not know how to spell the word. A typographical error, or typo, is the misspelling of a word because the user accidentally typed the word incorrectly. For example, the interlocutor could have accidentally pressed the wrong key or pressed keys out of order, swapping letters, as in *teh* for *the*. “The extent to which it is possible to determine whether spellings are ‘deliberate’ or mistaken is questionable” (Tagg, 2009, p. 133). Thus, these two features are grouped together.

Misspellings and typos have been found in online communication by a variety of researchers (such as Baron, 2008; Danet, 2001; Thurlow, 2003; Yongyan, 2000). Yongyan (2000, p. 33) believes these errors are “nothing to fuss over”; because cybermedia afford conversation so similar to face-to-face speech, it is not surprising that disfluencies should appear as they do in face-to-face speech, where they pass “more or less without comment.”

Silva (2010, p. 269) explains that errors are “not usually due to the lack of competence but because of performance demands.” For example, in N:N, synchronous communication where speed is of the essence, interlocutors may rush to compose a message and contribute it to the conversation before the tide of the conversation shifts away from the point they wish to make. In so doing, they may make errors.

Repairs

There are attempts to correct disfluencies in online communication (Cherny, 1999; Collister, 2008; Ferrara et al., 1991; Paolillo & Zelenkauskaitė, 2013). In Cherny’s (1999) corpus, interlocutors used programming expressions such as the search and replace command syntax—e.g., *s/<item to be found>/<item to replace item found>*—to indicate a repair. For instance, if someone mistyped the word *error* as *eror*, *s/he* might indicate a correction like so: *s/eror/error*. Collister’s (2008) examination of WoW chat and Wutiolarn and Attaprechakul’s (2010) examination of the game AuditionSEA revealed the use of asterisks to signify a repair, such as **error*.

Addressivity

Face-to-face cues, such as eye gaze and gesturing to someone, can help clarify who the intended recipient of a message is. In online communication, these cues are missing. To compensate, interlocutors may use a technique that Werry (1996) refers to as “addressivity” where the recipient’s userid is included at the outset of a message, followed by a colon. For example, in Werry’s examples, the interlocutor with the userid *boot* wishes to direct his

comment specifically to an interlocutor with the userid *Franck*. To do this, s/he types the message as follows: *frank: there's a girl*. Franck replies with *boot: where? where?*

Werry (1996, p. 52) explains that “addressivity is imperative on IRC, since the addressee’s attention must be recaptured anew with each utterance.” He believes the role of the listener can become passive and listeners do not have the opportunity to supply authentic (non-surrogate) minimal responses (e.g., *uh huh, mm hm*) to signal active engagement in the conversation. Addressivity may help interlocutors compensate “for the weakened link between sender and receiver” (Werry, 1996, p. 52).

Reduplication

Cherny (1999) discusses the duplication of certain words, usually without intervening spaces. She suggests that reduplication occurs because the interlocutor wishes to convey his/her sentiment more emphatically. Examples include *waveswaveswaves*, *nods nods*, and *nodditynodnods*.

Other word-creation processes

As with any examples of language, one might notice examples of new words being created through word-formation/creation processes such as affixation, compounding, and conversion. Affixation is a process where prefixes or suffixes are attached to words, such as attaching *-able* to *afford* resulting in *affordable* (Eble, 1996). Affixation does not include verb endings and plurals. Compounding combines two word bases to form a new word, such as *baseball* (Plag, 2003). In Yongyan’s (2000) analysis, unusual compounding was found where the whitespace between two words that would not ordinarily be compounded was

omitted. This special type of compounding will be referred to as space omission. Conversion is the process of shifting a word to a different grammatical class (e.g., part of speech) without derivational affixation, such as *hoof* for a way to move as in *hoofing it* (Crystal, 2008a; Eble, 1996; Rúa, 2007).

Cyberlanguage may include the use of these processes in new and unusual ways. Cherny (1999), Crystal (2006), Rúa (2007), and Yongyan (2000) provide several examples in their examination of cyberlanguage. Examples include:

<i>rehi</i>	affixation (Rúa, 2007), meaning <i>hi again</i>
<i>somuch, ihave</i>	compounding (Yongyan, 2000)
<i>eye</i>	used as a verb, conversion (Cherny, 1999)

Conclusion

These features, their definitions, examples, and the researchers who have shown evidence of them in online conversations are summarized in Table 2 below.

Table 2: Cyberlanguage features, definitions, examples, and sources.

Abbreviations			
Feature	Definition	Examples	Citations
Acronyms / initialisms	reducing a phrase to the initial letters of the words or syllables that words contain	<i>lol (laughing out loud)</i> <i>b/c (because)</i>	Baron (2003, 2008, 2010); Cherny (1999); Crystal (2006; 2008b); Danet (2001); Driscoll (2002); Ferrara et al. (1991); Hård af Segerstad (2002); Lewin and Donner (2002); Lindh (2009); Ling (2005); Liwei (2001); North (2007); Rúa (2007); Shortis (2007); Tagliamonte and Denis (2008); Thurlow (2003); Varnhagen et al. (2010); Werry (1996); Wutiolarn and Attaprechakul (2012)

Abbreviations cont.			
Feature	Definition	Examples	Citations
Shortenings	removing "meaningful" parts of a word (Crystal, 2008b, p. 50); including whole syllables or the removal of single vowels and consonants	<i>prob (probably)</i> <i>imedtly</i> <i>(immediately)</i>	Crystal (2008b); Danet (2001); Driscoll (2002); Ferrara et al. (1991); Hård af Segerstad (2002); Kadir, Maros, and Hamid (2012); Lindh (2009); Rúa (2007); Shortis (2007); Tagg (2009); Thurlow (2003); Werry (1996); Wutiolarn and Attaprechakul (2012)
Clippings	removal of the last letter (Crystal, 2008b, p. 45)	<i>goin (going)</i>	Crystal (2008b); Hård af Segerstad (2002); Shortis (2007); Tagg (2009); Thurlow (2003)
Single-letter forms	a single letter substituted for a word	<i>H (heroic)</i>	Baron (2010); Driscoll (2002); Hård af Segerstad (2002); Lindh (2009); Shortis (2007); Werry (1996)
Letter homophones	substituting a letter for a sound	<i>c (see)</i>	Baron (2008, 2010); Danet (2001); Hård af Segerstad (2002); Kadir et al. (2012); Lewin and Donner (2002); Rúa (2007); Shortis (2007); Tagg (2009); Thurlow (2003); Varnhagen et al. (2010); Werry (1996)
Number homophones	substituting a number for a sound	<i>gr8 (great)</i>	Crystal (2008b); Danet (2001); Kadir et al. (2012); Rúa (2007); Shortis (2007); Tagg (2009); Thurlow (2003); Varnhagen et al. (2010); Wutiolarn and Attaprechakul (2012)
Symbolic substitution	using non-alphabetical characters to signify a larger concept	?? (to signify confusion)	Cherny (1999); Danet (2001); Driscoll (2002); Ferrara et al. (1991); Hård af Segerstad (2002); Lindh (2009); Ling (2005); Rúa (2007); Wutiolarn and Attaprechakul (2012); Yongyan (2000)
Conjunctions and disjunctions	using a slash in place of "and" or "or" to signify an AND condition or an OR condition	<i>me/jay r</i> <i>coming 2</i> <i>do u want it</i> <i>now/later</i>	Christopherson (2013)
Punctuation omission	omitting traditionally-used punctuation	<i>dont (don't)</i>	Baron (2008, 2010); Cherny (1999); Cho (2010); Crystal (2006); Driscoll (2002); Hård af Segerstad (2002); Lewin and Donner (2002); Liwei (2001); Shortis (2007); Tagg (2009); Werry (1996); Yongyan (2000)
Non-standard use of lowercase	use of lowercase instead of typically-expected uppercase	<i>i like that</i>	Baron (2008); Cherny (1999); Cho (2010); Danet (2001); Driscoll (2002); Ferrara et al. (1991); Hård af Segerstad (2002); Lewin and Donner (2002); Ling (2005); Liwei (2001); Tagliamonte and Denis (2008); Varnhagen et al. (2010); Werry (1996); Wutiolarn and Attaprechakul (2012); Yongyan (2000)

Surrogate Prosodic Cues			
Feature	Definition	Examples	Citations
Onomatopoeic expression	sound effects; human vocalizations; noises	<i>muahahahaha!</i>	Cherny (1999); Crystal (2006); Danet (2001); Driscoll (2002); Kadir et al. (2012); Lindh (2009); Lewin and Donner (2002); Rúa (2007); Shortis (2007); Tagliamonte and Denis (2008); Thurlow (2003); Werry (1996)
Phonetic respellings	respelling a word to emphasize phonological aspects; often to simulate dialect/accent (Rúa, 2007); can include elisions where multiple words are joined to form a new word that emphasizes their elided pronunciation	<i>wot (what)</i> <i>tunez (tunes)</i> <i>whatcha (what are you)</i>	Baron (2008); Carter (2004); Cherny (1999); Crystal (2006; 2008b); Driscoll (2002); Ferrara et al. (1991); Hård af Segerstad (2002); Kadir et al. (2012); Lindh (2009); North (2007) Rúa (2007); Shortis (2007); Tagg (2009); Thurlow (2003); Varnhagen et al. (2010); Werry (1996); Wutiolarn and Attaprechakul (2012)
Offsetting punctuation	punctuation wrapped around a word; possibly for emphasis in the absence of bolding or italicizing functionality; or for decoration	<i>*that* is not good</i> <<<JOE>>>	Carter (2004); Cherny (1999); Crystal (2006); Danet (2001); Hård af Segerstad (2002); Lewin and Donner (2002); Lindh (2009); Liwei (2001); Werry (1996); Wutiolarn and Attaprechakul (2012)
All caps	capitalizing an entire word(s); possibly to indicate emphasis or to appear to yell	<i>THAT is not good</i>	Carter (2004); Cherny (1999); Cho (2010); Crystal (2006); Danet (2001); Danet et al. (1997); Hård af Segerstad (2002); Kadir et al. (2012); Lewin and Donner (2002); Werry (1996); Wutiolarn and Attaprechakul (2012); Yongyan (2000)
Surrogate Prosodic Cues cont.			
Feature	Definition	Examples	Citations
Letter duplication	duplicating letters; possibly to indicate elongated sounds	<i>Noooooooo</i>	Carter (2004); Cherny (1999); Cho (2010); Crystal (2006); Danet (2001); Danet et al. (1997); Hård af Segerstad (2002); Kadir et al. (2012); Lindh (2009); Shortis (2007); Tagg (2009); Tagliamonte and Denis (2008); Werry (1996); Wutiolarn and Attaprechakul (2012)
Punctuation duplication	duplicating punctuation; possibly to indicate tempo changes; emphasis; excitement; exaggeration	<i>not now..... im busy</i> <i>Nooooo!!!!!!!</i>	Baron (2008); Cherny (1999); Cho (2010); Crystal (2006); Danet (2001); Hård af Segerstad (2002); Lindh (2009); Ling (2005); Shortis (2007); Tagliamonte and Denis (2008); Thurlow (2003); Werry (1996); Yongyan (2000)

Surrogate Proxemic Cues			
Feature	Definition	Examples	Citations
Emoticons	“combinations of keyboard characters designed to show an emotional facial expression” (Crystal, 2006, p. 39)	:-((to indicate frowning)	Baron (2003, 2008, 2010); Carter (2004); Cherny (1999); Crystal (2006; 2008b); Danet (2001); Danet et al. (1997); Driscoll (2002); Hård af Segerstad (2002); Kadir et al. (2012); Lewin and Donner (2002); Ling (2005); Liwei (2001); Lo (2008); North (2007); Shortis (2007); Tagliamonte and Denis (2008); Thurlow (2003); Werry (1996); Wutiolarn and Attaprechakul (2012); Yongyan (2000)
Emotes	action-oriented text to indicate behavior or state of being	<runs away in terror> *chocolate grin*	Cherny (1999); Crystal (2006); Danet (2001); Danet et al. (1997); Hård af Segerstad (2002); Lewin and Donner (2002); Lindh (2009); Ling (2005); North (2007); Shortis (2007); Thurlow (2003); Werry (1996); Wilkins (1991); Wutiolarn and Attaprechakul (2012)
Pointing	punctuation forming an arrow pointing to the interlocutor’s username to indicate pointing to one’s self (Werry, 1996)	Bob <-- is dandy	Cherny (1999); Crystal (2006); Waskul and Douglass (1997); Werry (1996)
Pictograms	combinations of keyboard characters used to create a picture or graphic representation of a thing in the real world; what Danet (2001) would call “ASCII art”	(_/) (='.'=) (")__(")	Cherny (1999); Danet (2001); Shortis (2007); Werry (1996); Wilkins (1991); Wutiolarn and Attaprechakul (2012)
Errors and Repairs			
Feature	Definition	Examples	Citations
Typos and misspellings	failure to spell a word properly; possibly due to mistyping characters or lacking knowledge of the correct spelling	im stuffing my sace	Baron (2008, 2010); Crystal (2006); Danet (2001); Ferrara et al. (1991); Hård af Segerstad (2002); Rúa (2007); Tagg (2009); Thurlow (2003); Varnhagen et al. (2010); Wilkins (1991); Yongyan (2000)
Repairs	repairing a typo/misspelling	*face (the repair for sace)	Baron (2008); Cherny (1999); Collister (2008); Ferrara et al. (1991); Lindh (2009); Werry (1996); Wutiolarn and Attaprechakul (2012)

Other			
Feature	Definition	Examples	Citations
Addressivity	to avoid ambiguity and discontinuity in turn-taking; using the addressee's name and punctuation stylistically to indicate the intended addressee (Werry, 1996)	<i>joe: did you see my last msg?</i>	Werry (1996)
Reduplication	repetition of words or word roots in succession	<i>nodsnodsnods</i>	Cherny (1999); Crystal (2006); Rúa (2007); Wutiolarn and Attaprechakul (2012)
Affixation; compounding; blends; and other word-formation/word-creation techniques	<p>affixation – adding a prefix or suffix to a word (Eble, 1996)</p> <p>compound – combining two word bases to form a new word (Plag, 2003)</p> <p>conversion – when a word is shifted to a different part of speech or grammatical class without affixation (Crystal, 2008a; Eble, 1996; Rúa, 2007)</p> <p>blend – abbreviated forms that have been compounded (Eble; 1996)</p>	<p><i>webify</i></p> <p><i>ragequit</i></p> <p><i>nerf</i> (to make easier)</p> <p><i>fucktard</i></p>	Cherny (1999); Crystal (2006); Driscoll (2002); Hård af Segerstad (2002); Lindh (2009); Rúa (2007); Werry (1996); Yongyan (2000)

Word Creation

New words, and variations on old ones, are continually created as evidenced by the need to revise dictionaries to include them. New words are created by word-formation and other lexicalization processes. Crystal (2006, p. 71) explains that “the rate at which [cyberusers] have been coining new terms and introducing playful variations into established ones has no parallel in contemporary language use.” What follows is a discussion of how new words are typically formed so that word-creation strategies found in cyberlanguage may be evaluated, particularly for any departures from typical means of word creation, as these departures may be viewed as evidence of linguistic creativity. Then discussion will shift to

defining the related concept of productivity—the use of word-formation processes to create new words—and determining if it is synonymous or different from linguistic creativity. This will require further exploration of what it means for a word or utterance to exemplify creativity. Then, because creativity may manifest in language play, this section will conclude with a discussion of language play, and how cyberlanguage might provide evidence for linguistic creativity and play.

Word-formation

Word-formation refers to the creation of new words by compounding, conversion, combining forms, and affixation (Quirk, et al., 1985). Compounding is the process of “adding one base to another” such as combining *break* and *fast* to create the word *breakfast* (Quirk et al., 1985, p. 1520). Conversion is a derivational process whereby a word changes grammatical class without affixation (Crystal, 2008a; Quirk et al., 1985), such as the noun *hand* becoming the verb *hand* as in “hand it to me.” According to the American Heritage Dictionary online,²⁰ a combining form is “a modified form of an independent word that occurs only in combination with words, affixes, or other combining forms to form compounds or derivatives,” such as *electro-* combined with *magnet* to form *electromagnet*. Affixation is a process where prefixes and suffixes are attached to bases, such as attaching *-ment* to *establish* to create *establishment*.

Derivational processes, such as the ones discussed above, result in new words. Inflection does not, and so is typically not considered a word-formation process. Instead, inflection is a grammatical process of converting a word to its plural form (*house* to *houses*),

²⁰ See <http://ahdictionary.com/word/search.html?q=combining+form&submit.x=0&submit.y=0>

another verb tense (*work* to *worked*), or to demonstrate possession (*Laura's book*), but a new concept is not created (Crystal, 2008a). For instance, *house* and *houses* are fundamentally the same concept but *establish* (a verb) and *establishment* (a noun) are two different concepts, although related. With derivation, the grammatical class of the word may change—i.e., a verb may become a noun as in the case of *establish* and *establishment*—but not with inflection (Crystal, 2008a).

Some linguists, such as Bauer (1983) and Plag (2003) consider abbreviation to be a word-formation process. Bauer (1983), Plag (2003), Quirk et al. (1985), Zawada (2005), and others debate whether blends, back-formation, reduplication, and familiarity markers are examples of word-formation. Blends are the joining of two or more splinters (parts of words or clippings/shortenings) to form a new word (Lehrer, 2007). For instance, *brunch* combines the *br-* in *breakfast* with the *-unch* of *lunch*. Lipka (2007) believes blends are examples of word-formation. Back-formation is the derivation of a new word via the deletion of a suffix, as in *edit* from *editor*. Bauer (1983), Plag (2003), and Quirk et al. (1985) believe back-formation is a word-formation process. Reduplicatives are “compounds that have two or more constituents which are either identical or only slightly different,” such as *lovey dovey* (Quirk et al., 1985, p. 1579). Hård af Segerstad (2002) and Quirk et al. (1985) believe reduplication to be a word-formation process. Familiarity markers are words such as *sweetie*, *auntie*, and *Debs* (Quirk et al., 1985); Plag (2003) and Quirk et al. (1985) believe these are examples of word-formation.

There are other lexicalization processes that most scholars hesitate to officially deem as word-formation. These include borrowings—taking a word from one language and using it

in another language, such as using the Italian *pizza* in English—and figurative constructions such as allusions, metonymy, and metaphor.

Productivity and creativity

Productivity refers to the creation of new words and utterances through adherence to the rules of specific *languages*, such as French or German (Chomsky, 1966). For example, affixation is seen as a productive word formation process, and certain affixes are seen as more productive—capable of yielding more new words and forms of expression—than others. For example, *un-* is a more productive affix than *-th* because it can be attached to a larger number of words than *-th*, which is only attached to a very small set of words such as *length* and *width*. Furthermore, the degree to which an affix is productive may change over time—e.g., *-th* is rarely attached to modern words.

Creativity, according to Chomsky (1966) is a property of *language*²¹ as an act of communication—and has to do with how human communication is non-conditioned, as opposed to animal communication which is often regulated by stimulus response. Whereas animal communication is considered to be purely mechanistic, as a result of instinct, human communication is the “free expression of thought for appropriate response to new situations” (Chomsky, 1966, p. 13). Carter (2004, p. 78) claims that the Chomskyan notion of creativity is biological; it is “a statement about a genetically endowed capacity to exploit an underlying system.”

To Bauer (1983), productivity is rule-governed innovation while creativity is rule-bending innovation where the interlocutor extends the language an unpredictable way.

²¹ The distinction between *language* and *languages* comes from a personal email exchange with David Crystal on February 16, 2009.

Joining two combining forms such as *techno-* and *-crat* to form *technocrat* follows a convention for joining words or combining forms that end in vowel sounds with the combining form *-crat*, to create terms for specific types of rulers/leaders, such as *bureaucrat* and *autocrat*. This is rule-governed innovation. “Lexical creativity arises when old devices are used in new ways” (Rúa, 2007, p. 157), such as attaching “existing affixes to unusual or unorthodox bases” as in the case of attaching the prefix *re-* to *hi* to form *rehi* (Rúa, 2007, p. 147). Affixes are not typically attached to greetings, such as *hi* or *hello*. *Rehi* is an example of rule-bending innovation.

Thus, productivity is a form of creativity in that new forms are created through adherence to a language’s rules about word-formation, but it is a mechanistic process that presumes some innate human ability to form new, stimulus-free utterances. This is the Chomskyan notion of creativity. Alternatively, linguists, such as Bauer, speak of a different form of creativity that is not mechanistic and biologically-driven. It is a form of innovation where new forms are created by exploiting and bending a language’s rules. Linguistic creativity is further specified as the creation of novel forms as a way to fulfill some social and communicative purpose—including overcoming conceptual gaps—within a specific context with specific individuals.

“Most approaches to creativity relate it to novelty” (North, 2007, p. 539). Interlocutors purposefully create new forms because it is enjoyable, indexes personal identity, and demonstrates one’s sense of belonging to a group (Carter, 2004). “Creativity functions to give pleasure, to establish both harmony and convergences as well as disruption and critique, to express identities and to evoke alternative fictional worlds which are recreational and which recreate the familiar world in new ways” (Carter, 2004, p. 82).

Interlocutors may also choose to create new forms to demonstrate politeness, avoid sensitive issues, or enhance one's status within a group (Zawada, 2005). Creativity is also risky because there is always the possibility that a new form will fall flat in some way, and an unsuccessful performance can result in embarrassment; a successful one can create “accord, intimacy, involvement, affect” and can be “schema-refreshing” (Carter, 2004, p. 110).

Novel expressions may also be required to overcome problems within a “conceptual space” (Carter, 2004, p. 36). Creativity may be employed to resolve linguistic gaps or conflicts within a communication situation, as in the case of creating a new term for a new concept that hitherto had no sufficient label for it (Howden, 1984; Quirk et al., 1985; Zawada, 2005). This may be done by combining elements in new ways, such as adding affixes to unorthodox bases, compounding terms that have not previously been combined, or in more rare cases, creating something from nothing (e.g., coinage).

Creativity is context-dependent (Carter, 2004; North, 2007). It is dynamic and emergent, “relative to the values, beliefs and judgments formed within and according to the needs of different social groups, communities and cultural systems” (Carter, 2004, p. 82). It requires “‘insider’ recognition and acceptance” (Carter, 2004, p. 140). “Creative processes and creative thought have to be adaptive and to be fitted to a changing environment and existing social conventions” (Carter, 2004, p. 41). “Creativity results in changes to domains or in the establishment of new domains” (Carter, 2004, p.48).

Creativity is also specific to the individual (Quirk et al., 1985). It is the result of dialogic interaction among individuals (Carter, 2004). A creative individual is someone who possesses the ability “to think laterally and innovatively, especially for purposes of problem-solving and changing accepted ways of seeing and understanding” (Carter, 2004, p. 41). In

speaking of himself, Picasso (as cited in Picasso & Sabartés, 1946)—²²generally thought to be a creative genius—explained that if he were to adhere to grammatical rules that “have nothing to do with me, whatever is personal in my writings would be lost in a grammar which I have not assimilated. I would prefer to invent a grammar of my own than to bind myself to rules which do not belong to me.” “Competent users of a language have an extended language repertoire, and when new situations arise, they create new appropriate language varieties out of existing language varieties” (Ferrara et al., 1991). “Every individual creative act of every speaker therefore, has the potential to change the language, in the sense of add-ons, growth and development (e.g. in the vocabulary), as well as in the modification of the system” (Zawada, 2005, p. 49).

Language play

Language play is very much a linguistically creative process. Carl Jung (1971, p. 200) explains that “the creation of something new is not accomplished by the intellect but by the play instinct.” As with linguistic creativity, play is about breaking the rules (Crystal, 1998). When some linguistic feature—letters, sounds, words, word parts, phrases, sentences—is manipulated—i.e., made to do something it would not normally do—for the purpose of enjoyment, this is play (Crystal, 1998).

Language play can include contrasting tones of voice, sound play, mock regional tones, jokes that make plays on words, manipulation of letters, deviant forms of monologue and dialogue, word games like crossword puzzles, word twisting like puns, nonsense scat singing, rhyme, reduplication, morphological changes such as adding endings to nouns (*fishy*,

²² See p. 119 in Picasso, P., & Sabartés, J. (1946). *Paintings and drawings of Picasso, with a critical survey*. Paris: Braun & cie.

snakey), nonsense names (*Mr Higglety Pigglety*), and code languages like Pig Latin (Crystal, 1996).

Like creativity, play involves metalinguistic awareness of the structure of language so that the norm can be reshaped in new ways (Crystal, 1996; Danet et al., 1997). Words are treated as objects or toys to be played with (Danet et al., 1997). Language play is about “upping the ante,” stretching conventions, as in taking *IMHO* (*in my humble opinion*) up a notch to *IMHBCO* (*in my humble but correct opinion*) (Crystal, 2008b, p. 53).

Language play is important personally (adding to quality of life), socially (signifying group bonds), educationally (improving language learning for children), and creatively (as a means of self-expression for a variety of domains). Poets, advertisers, comedians, and more all engage in language play for creative expression.

Conclusion

Crystal (2008b, p. 27) explains that many linguistic processes used in creating cyberlanguage are “centuries old.” They have been seen in cartoons, advertisements, and poetry; however, “what is new, though, is their simultaneous and worldwide usage” (Silva, 2010, p. 267). Interlocutors draw on what they already know about language to satisfy the demands of new and changing communication situations. Overcoming and working with the constraints and affordances of new media to refashion language so that it suits the context exemplifies creative problem solving.

Crystal (2006, p. 71) believes that “a strong personal, creative spirit imbues Netspeak, as an emerging variety.” Online language may be the “latest manifestation of the human ability to be linguistically creative and to adapt language to suit the demands of diverse

settings” (Crystal, 2008b, p. 175). It may very well be resounding evidence of human ability to create and play with language, something that makes humans very special—“homo loquens at its best” (Crystal, 2006, p. 276).

Online, the “distinction between text and context becomes blurred”; “context is itself textually-constructed” (North, 2007, p. 540). Because cybermedia introduce new and changing communication opportunities that are socially co-constructed and that present new communication puzzles to be solved, they are fertile ground for creative expression (Carter, 2004; Rúa, 2007). “Internet users are continually searching for vocabulary to describe their experiences, to capture the character of the electronic world, and to overcome the communicative limitations of its technology” (Crystal, 2006, p. 71). This “new world of technology” has led to “almost endemic, and to a certain degree essential” coinage (Quirk et al., 1985, p. 1535). “The medium has provided an impulse towards new text types and new forms of creative interaction, in which a new interface has been created between spoken and written language” (Carter, 2004, p. 190).

Carter (2004, p. 193) considers the “grapho-phonemic manipulations of the language system”—such as capitalization or duplication of letters to indicate emphasis and vocal quality—found in online communication to be creative. Rúa (2007) deems creative the use of many of the features outlined in the earlier section “Cyberlanguage and Its Characteristics.” Some of these features are examples of the word-formation processes outlined above; some represent other word-creation strategies that involve other forms of creative manipulation, such as typographical and orthographical variation.

Furthermore, “wordplay is ubiquitous” in cybermedia (Crystal, 2006, p. 171). Danet et al. (1997) believe that online language is inherently playful because the object that

facilitates the communication—the computer—necessarily invites experimentation and bricolage—i.e., creativity. The ephemerality, speed, interactivity, and “freedom from the tyranny of materials” that a computerized, virtual environment offers is what fosters playfulness in computer-mediated communication (Danet et al., 1997, *An Inherently Playful Medium*, para. 2). With cybermedia, interlocutors may “invoke the frame of ‘make-believe’” (Danet et al., 1997, *An Inherently Playful Medium* section, para. 1). Identities are masked that may free interlocutors to be “other than ‘themselves’” (Danet et al., 1997, *The Masking of Identity*, para. 1) so they may “experiment with different forms of communication and self-representation” (Reid, 1991, *Computer-Mediated Communication* section, para. 7). Communication in cybermedia may, therefore, become performative (Danet, 2001; Danet et al., 1997). The stage is simply the range of typographical choices a keyboard offers, and the script is what the interlocutors make of it, moment-to-moment, with their creative manipulation of language.

Conclusion

As Table 2 shows, there was a flurry of cyberlanguage research that specifically examined lexis in the 1990s up to the early-to-mid 2000s. These studies made assertions about the influence these media (e.g., forums, email, SMS, IM, and chat) and their characteristics (e.g., synchronicity, participant scale, message permanence, privacy, anonymity, message length restrictions, and compositional and viewing ease) have on language production. For example, media with message length restrictions are thought to encourage more abbreviated forms.

Then, in the 1990s and 2000s, research into computer-mediated communication began to explore other aspects of online communication, such as turn-taking, cohesion, politeness, and gender and cultural differences (Herring, Stein, & Virtanen, 2013). It is possible that research began to shift away from lexical analyses because the community believed it had exhausted this line of research; however, this is not clear given the unanswered questions that remain. For example, to this researcher's knowledge, no one has yet examined language across a variety of social media to explore the validity of suppositions made about their effect on language production, which is the goal of this study.

Most of these early studies (as well as the few recent studies) focused on a single medium in isolation and thus could not make comparisons of language across media, nor provide a broader, more comprehensive view of cyberlanguage use. For example, Lewin and Donner (2002) and Kadir et al. (2012) examined discussion forums. Cho (2010), Danet (2001), and Yongyan (2000) examined email. Baron (2008), Crystal (2008b), Ling (2005), Tagg (2009), and Thurlow (2003) examined SMS. Baron (2008), Ferrara et al., (1991), Tagliamonte and Denis (2008), and Varnhagen et al. (2010) examined IM or an IM-like medium. Cherny (1999), Danet (2001), Driscoll (2002), Werry (1996), and Wutiolarn and Attaprechakul (2012) examined chat.

Some work has attempted comparison of language across media via ex post facto analysis of different study findings that may have asked different research questions and used different methods on corpora of different sizes. For example, Baron (2008) compared findings from two studies of SMS and IM, and Hård af Segerstad (2002) compared findings from separate studies of email, SMS, IM, and chat.

Furthermore, many researchers have defined the cyberlanguage features differently. Some provided feature frequencies, such as Baron (2008), and some did not. Some grouped features into categories and then reported frequencies at the aggregate level (e.g., Lewin & Donner, 2002). Without a consistent set of features, feature definitions, and individual feature counts per medium to guide a comparison of language across media, it is difficult to know how feature use may change depending upon the situation. This makes it difficult to evaluate assertions made about cybermedia's influence on language production, or to provide a comprehensive definition and description of cyberlanguage.

Specific goals of this study are:

- to compare language—specifically cyberlanguage feature frequencies—across media, specifically forums, email, SMS, IM, and chat,
- to compare language—specifically cyberlanguage feature frequencies—across situational factors (or genre factors), such as topic and purpose,
- so as to test assertions made about media and situational influences,
- and to provide a more comprehensive description of cyberlanguage as it manifests across media.

Thus, the study described in this document differs from earlier studies in its use of a larger corpus that spans multiple types of media and is framed by a consistent set of research questions, a consistent catalogue of features and their definitions derived from a review of the literature, and a consistent methodological approach. Crystal (2006) explains that a systematic, empirical observation of this sort has yet to be pursued, and that no corpus of the sort used in this study has been created, and that without this, a great deal of what we know about online language will be subjective. One would assume that the length of time required

for such an analysis may be cause for why this sort of study has not been conducted. However, the work is needed if we are to better understand how technology influences communication and information behavior, and how users respond and adapt to technology and technological change. Crystal (2006, p. 275) believes it is possible that “we may one day communicate with each other far more via computer mediation than in direct interaction.”

The readiness with which people do adapt language to meet the needs of new situations, which is at the heart of linguistic evolution ... is going to be fully exploited in the next few decades, with the emergence of yet more sophisticated forms of digitally mediated communication. (Crystal, 2006, p. 257)

Online language is “a development of millennial significance” and a “new medium of linguistic communication does not arrive very often, in the history of the race” (Crystal, 2006, p. 272). Cyberlanguage will continue to grow and change and some baseline description of it across situations is needed to serve as a springboard for deepening our understanding of how technology influences and shapes the linguistic contours of our interactions, and understanding how people respond and adapt to technological change and how such adaptations are evidence of linguistic creativity. The achievement of these goals is addressed by the following guiding questions:

RQ 1: What cyberlanguage features are common across online, conversational media and genre situations?

RQ 2: What cyberlanguage features differ between media and genre situations and how do they differ?

These two research questions focus on media and genre as influencing factors by testing assertions made in prior research about their impact on language production. Specifically they ask if there are significant similarities or differences in the types and frequencies of linguistic features used in different media and genre situations. For example, subquestions might include: Do acronyms appear more frequently in synchronous media (e.g., chat, IM) or asynchronous media (e.g., forums, text messaging, listservs); do emoticons appear more frequently in situations where the main topic of discussion is gaming or in situations where non-gaming topics are discussed? The answers to these questions will constitute a description of cyberlanguage as it appears in online situations, and will result in a lexicon of terms and a grammar of features that may aid information professionals in designing tools and techniques to better search and mine these media for the potentially valuable information they may contain.

The third, more exploratory research question seeks to augment understanding of the user side of this equation by uncovering examples that represent the ability to creatively and innovatively modify interactions when faced with change in the environment:

RQ 3: Are there examples of linguistic creativity that may serve as evidence of an interlocutor's ability to respond and adapt to technological change in innovative ways? If so, what are some examples?

Methods

Introduction

To answer the research questions detailed in the Literature Review section, a corpus was created that contained conversation from online, conversational media—specifically forums, email lists, SMS, IM, and chat. The corpus was balanced, with roughly equal proportions of words from each of these five media, and as equal as possible proportions of words spanning certain topics (gaming, technology, other) and certain purposes (serious, recreational).

This is an exploratory, lexical analysis of the individual terms contained in this corpus. They were examined for the presence of the 25 linguistic features shown in Table 2: Cyberlanguage features. Because the unit of analysis is the word (or individual term), a working definition of *word* is required to guide the analysis. A rather rudimentary definition of *word* is a set of characters preceded and followed by whitespace, but this definition is infrequently used in traditional language studies because (a) many words cross whitespace boundaries, (b) contractions eliminate intervening whitespace but are rarely treated as single words, and (c) punctuation that may be attached to a word is thought to be non-word information that should be removed prior to analysis. *Bird cage* is an example of (a). Because of *bird* and *cage*'s frequent co-occurrence, they may be considered to be one unit—i.e., one word, a compound. *Don't* is an example of (b). Contractions eliminate whitespace and letters of two words, and most language studies would treat *don't* as two distinct words: *do* and *not*.

Quotation marks, periods, exclamation points, and other punctuation that is appended to a set of characters is an example of (c) because punctuation marks are not typically considered to carry any semantic weight. A period has no definition in a dictionary, for example. It is there to mark the ending of sentences, and so has a more presentation-oriented, discourse-formatting function rather than a semantic one. Thus most lexical studies remove punctuation before analysis.

The precise definition of a word used in any study depends in part on the purpose of the study. Cyberlanguage deviates quite a bit from standard or general English, and these deviations are of interest in this study; so preserving them is important. For example, some cyberlanguage terms consist primarily of punctuation, as in the case of emoticons (e.g., :-O). The colon and dash in this example do carry semantic weight. Combined with the capital *O* they portray an interlocutor's emotional response to conversation. Removing any of the punctuation in an entity like :-O would result in a loss of valuable communication information.

Additionally, most natural language processing tools, trained on general English corpora, are not equipped to process cyberlanguage terms. For example, in cyberlanguage, punctuation is often omitted from contractions, so most preprocessors that rely on the presence of an apostrophe to divide the contraction into two separate words would fail to make this division.

So given the departures cyberlanguage takes from general English, commonly used definitions of *word* in most linguistic analyses may be unsuitable for cyberlanguage analysis. Thus, the more basic definition of *word*, *term*, or *lexical item* will be used and further specified as a set of characters, punctuation, and numbers preceded and followed by

whitespace.²³ This definition makes word-counting “very reliable” (Kilgarriff, 1997, p. 233).

It also allowed the researcher to treat all units, including those that deviate from general

English as words of interest, e.g.,:

<i>^ ^</i> <i>_</i>	an emoticon
<i>l8r</i>	<i>later</i>
<i>ARGHHHHH!!!!!!!</i>	emphasized onomatopoeic expression

The rest of this section provides further details about the creation of the corpus and the analysis methods. It begins by discussing the principles of corpus design, including exploration of the concepts of balancing and representativeness as they are defined within the domain of corpus linguistics. A discussion of specific criteria used to balance the corpus and ensure its representativeness follows. Then decisions about corpus size are outlined. The corpus is described in detail including sampling and collection methods used to create it. Then analysis methods are outlined, including a description of the classification process, the inter-coder reliability test, and the statistical tests used.

Corpus construction and analysis comprised eight steps:

1. determined criteria for corpus balancing based on a review of the literature, and determined viable sources for texts;
2. sampled texts from selected sources;
3. cleaned texts of non-conversational text;
4. created a list of all terms and their frequencies in the corpus;
5. from this list, removed standard/general English terms;

²³ This definition is also used by Silva (2010) in analysis of chat.

6. conducted an inter-coder reliability test by enlisting the help of an outside coder to classify 5% of the remaining terms by the features listed in Table 2;
7. analyzed all terms by classifying them according to linguistic feature, determining their meanings, and recording examples of creative word-formation processes that illustrate an interlocutor's ability to bend and co-opt the rules of English to suit new situations; and
8. compared feature frequencies across media and genre factor using chi-square tests.

Corpus Design

Having a data set that is representative of the population being studied is important for making generalizations about that population. With corpora—a type of data set—representativeness is typically defined differently than in studies that examine human populations. This is because, with language, “it is difficult to define what the total population is, and...the population is continually growing” (Atkins & Rundell, 2008, p. 64). With human populations, however, census, tax, and voting records can be used to define a population and estimate its size (at least in the U.S.); researchers can also obtain, with greater certainty, a sample that is representative via random sampling. Kilgarriff (2005) explains that language, however, is never, ever random; so random sampling is inappropriate for language studies. “Words do not occur according to the laws of chance” (Sinclair, 2005a, p. 11).

Proportional sampling is also problematic in language studies because “there are no such things as ‘correct proportions’ of components of an unlimited population” (Sinclair, 2005a, p. 2). “A key aspect of corpus design for most studies, then, is including the range of linguistic variation that exists in language, not the proportion of variation” (Biber, Conrad, &

Reppen, 1998, pp. 247-248). Thus, representativeness is concerned with “reflecting the diversity of the target language” (Atkins & Rundell, 2008, p. 66); or as Gries (2009, p. 1231) explains: “the different parts of the linguistic variety I’m interested in are all manifested in the corpus.” Representativeness is achieved through a process of balancing where the research includes “texts which collectively cover the full repertoire of ways in which people use the language” (Atkins & Rundell, 2008, p. 66).

A stratified approach is taken where the corpus builder “catalog[ues] the different categories of texts that exist in a language and sampl[es] each of them” (Biber et al., 1998, p. 248). These stratifications—or criteria categories—reflect the functions and situations of language (Atkins & Rundell, 2008). Although there is no “universally agreed classification of text-types” (Atkins & Rundell, 2008), researchers do make some recommendations, many of which overlap with the components of genre as defined earlier in this paper. Some of these recommendations include:

- Purpose (Biber, 2008; Cabré, 1999),
- Topic or subject matter (Atkins & Rundell, 2008; Biber, 2008; Cabré, 1999),
- User community including scale (i.e., number of interlocutors) (Biber, 2008; Cabré, 1999),
- Communicative situations, settings, domains (Biber, 2008; Cabré, 1999; Sinclair, 2005a),
- Monolingual, bilingual, multilingual (Atkins & Rundell, 2008; Sinclair, 2005a),
- Timing, dates; synchronic or diachronic (Atkins & Rundell, 2008; Sinclair, 2005a),
- Mode: spoken, written, electronic, some combination (Atkins & Rundell, 2008; Biber, 2008; Sinclair, 2005a),

- Format / text types (Atkins & Rundell, 2008; Biber, 2008; Sinclair, 2005a), and
- Location (Sinclair, 2005a).

According to Sinclair (2005a), once criteria categories are defined, sources for text types that satisfy these criteria can be identified. Target sizes for each category are determined by the criterion's importance in answering the research questions, as well as the availability of texts and feasibility in obtaining them. "Ideally not only should all parts of which a variety consists be sampled into the corpus but also that the proportion with which a particular part is represented in a corpus should reflect the proportion the part makes up in this variety and/or the importance of the part in this variety" (Gries, 2009, p. 1231). Ideally, categories that are more important for answering the research questions would contain more text than categories that are less important to answering the research questions. However, in the case of each category being equally important for comparing language, roughly equal proportions would be ideal. "Strata can be fully represented (100 % sampling) in the proportions desired, rather than depending on random selection techniques" (Biber, 2008, p. 65). However, once the proportions are decided for each category, texts that satisfy those categories could be selected at random (Atkins & Rundell, 2008).

During corpus building, attention should be given to avoiding "rogue" documents that can lead to a "skewed" corpus (Atkins & Rundell, 2008, p. 63). A rogue document is one that has the potential to over represent a particular feature because it may contain vocabulary that does not appear in other parts of the corpus. For instance, suppose a corpus was to be designed to include texts from various fiction genres, and suppose that 10 texts were collected from classical fiction, 15 from mystery, 12 from fantasy, and one from science fiction. Then suppose the science fiction text discussed subatomic particles at length. The

corpus might then appear to have more vocabulary and grammatical constructions pertaining to physics than it would really have if the science fiction category had contained a higher number of texts featuring a richer distribution of topics discussed.

Balancing criteria

For the study discussed in this dissertation, there are two main divisions of criteria: criteria pertaining to medium—hereafter referred to as media factors—and criteria pertaining to the communication situation or genre—hereafter referred to as genre factors. Texts were also selected based on some source and text-quality requirements largely for reasons pertaining to the study’s scope and focus (e.g., collecting only texts that include conversation per the definition of conversation²⁴ discussed in the Cybermedia chapter) and feasibility (e.g., collecting only English texts because the researcher does not read other languages).

Media factors include the type of medium (e.g., forums, email) as well as the media characteristics (e.g., synchronicity, participant scale, message length restrictions) that were discussed in more detail in the Cybermedia chapter. The types of media selected for analysis include forums, email in the form of email lists,²⁵ SMS, IM, and chat. As was stated in the Literature Review, these media were targeted for analysis because (a) they are used for the purpose of conversation, and (b) they have been studied in previous language studies, which

²⁴ Recall the definition of *conversation* is a dynamic, back-and-forth flow of comments and responses—a repartee, a discussion—by two or more interlocutors where thoughts tend to be shared in an extemporaneous manner with less planning and editing than one might see in more formal texts such as scholarly articles, brochures, news articles, novels, etc. The text is necessarily dialogic, not monologic. In other words, the text is not for the purposes of one-sided broadcasting of thoughts and ideas. Conversation emerges, unfolding organically over a period of time and having an unpredictable focus, rather than being created holistically with a predetermined focus as one might find in scholarly articles, news articles, novels, etc.

²⁵ Email lists, instead of personal email, were selected because messages were publicly available through the Google Groups website.

makes it possible to test prior assertions about the effects these media and their characteristics have on language production.

In the Literature Review, genres were defined as text types that have a specific purpose(s) and are used by a particular group of interlocutors who follow certain norms and have certain expectations in a particular situation. These text types usually contain specific content (e.g., topics) and are formatted in specific ways. Two of these facets of genre—topic and purpose—comprise the genre factors category. Interlocutor and community-specific information, such as norms and expectations, were not used as genre factors in this study, because obtaining personally identifying information about interlocutors creates risk to privacy and obtaining permission from the thousands of interlocutors whose comments appear in the corpus can be time-consuming and difficult, if even possible. Format was also not selected for inclusion in the genre factors category because in these cybermedia, interlocutors have little, if any, control over message presentation. Usually the medium controls all or most of how messages are formatted. Purpose and topic, however, presented fewer barriers, and along with the particular source—which could be seen as representing a particular situation—they intertwine to suggest a particular social and communication environment.

Three topics were selected for comparison: Gaming, Technology, and Other. Cherny (1999) and Crystal (2006) explain that online language is an invention by early Internet programmers (or hackers), and was adopted and grew in communication situations like early MUDs and MOOs—precursors to today’s MMOGs. Therefore, it would be worthwhile to compare language samples from media that foster conversation about gaming and technology topics, and then compare that to language samples from media that foster conversation about

other kinds of topics to determine if more cyberlanguage features are present in media that focus on gaming and technology topics. These three topics are specifically defined as follows:

- Gaming - Discussion is largely about computer and/or console games. Subtopics may include game play, the game world (e.g., game items, locations on the game map), obtaining help with and completing game tasks, coordinating game activities, and socializing in the game world.
- Technology - Discussion is largely about computer technology. Subtopics may include hardware or software configuration, user-interface problems, operating systems, programming, help and how-to information, networking issues, data storage issues, interoperability, etc.
- Gaming-technology - This topic is a specialized form of the Technology topic. Discussion is largely about computer technology as it relates to a game. Subtopics include hardware specifications for running a game, configuration of game software, technical support issues related to game performance, programming and coding help if the game allows gamers to modify some or all of it, networking and latency issues that affect game play, etc.
- Other - Discussion that is not related to gaming or technology. This could include topics related to health and wellness, other kinds of hobbies beside computer gaming such as gardening, activities around the home, music, art, television shows, movies, politics, business concerns, volunteer work, school, etc.

Two purposes were selected for comparison: Serious and Recreational/Leisure-oriented. Prior studies have examined language from media intended for serious, work, or

school-related communication as well as from media for recreational communication. When considering the results from these various studies (referenced in Table 2), it appears that more cyberlanguage features are discussed in the studies that examine recreational or non-work/school-related communication, such as Cherny (1999), Danet et al. (1997), Lewin and Donner (2002), Thurlow (2003), and Werry (1996). Thus, comparing language samples from media intended for serious communication with samples from media intended for recreational communication may help determine if purpose has an effect on language production in cybermedia. These two main purposes are defined as follows:

- Serious - The source is largely intended to support communication that is work-related, school-related, or for some other serious, non-recreational purpose.

Communication may be obligatory or compulsory in some way. It may be for some solemn intention. It is not superficial or light, and not related to amusement.

- Recreational/Leisure-oriented (non-serious) - The source is largely intended to support non-work-related, non-school-related, or recreational and leisure-oriented communication. Subpurposes could include to communicate about a hobby or game, to socialize and make friends, to discuss the events of one's day, etc. It is un-coerced, voluntary communication that usually takes place in one's free time. It is not for work but for pleasure, enjoyment, and diversion.

In addition to these media and genre factors, texts and their sources were selected based on their ability to satisfy some additional requirements that pertain to study scope and feasibility. First, texts were drawn only from sources that support conversation—as defined in the Cybermedia chapter—because most prior research on cyberlanguage has tended to focus on conversation—not narratives, stories, news articles, journals, and the like.

Replicating that focus permits evaluation of assertions made about the effects cybermedia may have on language production. Thus, texts were not drawn from blogs, Twitter, or Facebook wall posts, because these tend to be monologic, often one-off status updates, sometimes with intermittent and therefore unreliable bursts of true conversation.

Furthermore, texts that contain high incidence of programming code, advertisements and solicitations, magazine or other publication reprints, and bot²⁶ communication were not collected, because these are not conversation as defined in this paper. One might make the case that bot utterances are a form of conversation, but no human being is actively responding to the conversation; instead, utterances are automated by a computer program and have little to no relationship to the flow of human conversation into which they are inserted.

Second, texts were drawn only from sources that attract sufficient interlocutor participation. *Sufficient* is defined as frequently occurring comment-response pairs and/or utterances in close chronological succession to one another. In other words, a discussion forum thread was not sampled if there were no replies to the original posting. A chat room was not sampled if people sat idly by, contributing no or very few utterances over several minutes or an hour. Sufficient participation is also important in lieu of the ability to obtain what would traditionally be considered a complete text. Sinclair (2005a) encourages corpora developers to obtain complete texts, which he explains is more important than having categories with perfectly equal word counts. In cybermedia, there are often no clear starting and end points to a conversation, making obtaining “complete” texts difficult if not impossible. For example, forums group messages into threads usually centered on a particular topic; and although the original post could be considered a clear beginning to the

²⁶ Bots are “software applications that run automated tasks over the Internet” (http://en.wikipedia.org/wiki/Internet_bot). An example is a sexbot that is programmed to continually post text containing a URL to another website that allows one to purchase sexual favors.

conversation, interlocutors may continue to post responses indefinitely. Even more complicated is chat. If the chat room is persistently available, conversation may never conclude. For example, World of Warcraft is a persistent virtual world, so chat is always available and always active. Without clear starting and end points, it is difficult to say any cybermedia text—or message or thread—is complete. Thus, the requirement of sufficient participation can act as an approximation for “completeness” in that it ensures that texts are packed “full” with conversation.

And finally, only predominantly English texts were collected because the researcher does not read other languages.

Corpus size

In addition to determining *what* the corpus should include, decisions were required about *how much* the corpus should include. However, there are “no reliable guidelines as to what quantity of text represents a representative corpus” (Sager, 1990, p. 130). Sinclair (1991, p. 18) suggests that “a corpus should be as large as possible, and should keep on growing.”

Biber et al. (1998, p. 249) suggest that “enough texts must be included in each [criterion] category to encompass variation across speakers or authors.” But how much is enough? In Biber’s work with the Lancaster/Oslo-Bergen (LOB) corpus, he concluded that 10 texts per category were sufficient and that “counts are relatively stable across 1,000-word samples from a text” (Biber et al., 1998, p. 249).

Well-known corpora intended for the study of general language, such as the Brown corpus and the LOB corpus, started with 500 text samples of approximately 2,000 words

each, with 15 categories. Each was approximately 1 million words. Sinclair's Bank of English (Collins Cobuild) corpus had 450 million words and is still growing. The British National Corpus (BNC) had more than 4,000 text samples of approximately 100 million words. The lexicographers for the Oxford English Dictionary used 5 million excerpts of English (Sinclair, 1991).

Studies that have examined cyberlanguage are comparatively smaller. Table 3 lists the dimensions of some of the corpora used in the studies cited in this paper. Studies that did not report size or did not appear to examine most or all of the vocabulary are not listed. Not all these studies performed the kind of lexical analysis described in this dissertation where each word and its various usages in the corpus were examined, but most examined various aspects of the bulk of the vocabulary.

Table 3: Dimensions of cyberlanguage corpora.

Researcher	Medium	Messages	Words	Other Size Reports
Kadir, Maros, & Hamid (2012)	Forums	110		Collected a little over 3 months
Lewin and Donner (2002)	Forums	200		5 different forums
Cho (2010)	Email	197	16,569	
Danet (2001)	Email	20		
Gains (1999)	Email	119		
Hård af Segerstad (2002)	Email	183	11,660	
Waldvogel (2007)	Email	515		
Baron (2008), Baron and Ling (2011), Ling and Baron (2007) (reporting on analyses of the same corpus)	SMS	191	1,473	
Bieswanger (2007) The size shown here is only for Bieswanger's English corpus.	SMS	201	1,120	
Hård af Segerstad (2002)	SMS	1,152	17,024	
Ling (2005)	SMS	867		
Ling & Baron (2007)	SMS	191	1,473	
Tagg (2009)	SMS	11,067	190,516	
Thurlow (2003)	SMS	544		
Baron (2008), Baron (2010) (extensions of the study reported in Ling & Baron, 2007)	IM	2,185	11,718	23 conversations

Researcher	Medium	Messages	Words	Other Size Reports
Hård af Segerstad (2002)	IM	8,255		
Ling and Baron (2007)	IM	191	1,146	
Cherny (1999)	Chat			25 MB of logs
Collister (2008)	Chat			5-10 hours per week over 6 months
Danet et al. (1997)	Chat			1.5 hours logged
Driscoll (2002)	Chat			75 log files collected over 6 months
Hård af Segerstad (2002)	Chat	44,380	410,355	120 hours logged on one channel
Silva (2010)	Chat	10,685		90 minutes: nine 10-minute sessions
Werry (1996)	Chat			2 ten minute logs

Many of the studies listed in the above table do not report word counts, and they use various means to characterize the size of the corpus, such as reporting numbers of messages, numbers of hours logged, or megabytes of storage. Unfortunately, the lack of consistency in size reports makes it difficult to use these studies as a guide for determining an appropriate corpus size. However, as Table 3 shows, most of the studies that did report word counts used a relatively small number of words—fewer than 20,000. A couple used significantly more—Tagg and Hård af Segerstad—but neither of these reported examining each word individually. This would suggest that corpora consisting of fewer than 20,000 terms are most common for studies of a single medium that examine most or all of the terms in the corpus individually. Thus, for a study of language across five types of media that also spans multiple topics and purposes, as this dissertation study does, a corpus five times larger than the norm for these single media studies is probably more appropriate. Determining the exact figure, however, also requires that some attention be given to the type of methods used to analyze the corpus.

A *manual* examination of all cyberlanguage terms in the corpus was planned because automatic means—discussed in more detail later in this section—would be inappropriate for

fully answering the research questions. This is largely due to the non-standard, novel nature of the language, which most natural language processing tools, such as WordSmith or Natural Language Toolkit (NLTK), are ill-equipped to handle, particularly in regard to achieving the goals of this study. However, without the aid of these tools, analysis of thousands of unique word types would prove incredibly time-consuming. Therefore, to complete the dissertation in a reasonable timeframe, it was important to obtain a more manageable volume of text than what well-known corpora (such as Brown and LOB) include. Additionally, the corpus used for this study focuses on a special subset of language. Its creation was not intended for analysis of general language, and as such, probably does not need to be quite as large as Brown or LOB, at least initially; but it is hoped that this corpus will continue to grow in the future. Thus, a corpus size of approximately 150,000 words—30,000 per medium—was selected. This amount exceeds what is typically collected per medium for studies of cyberlanguage, yet is not overly voluminous so as to prolong analysis beyond a reasonable timeframe.

Corpus Creation

Sources were selected based on the availability of texts and on the criteria outlined in the Corpus Design chapter. Sources

- providing opportunities for forum, email, SMS, IM, and chat-based conversation,
- focused on Gaming, Technology, and Other topics,
- used for Serious and Recreational/Leisure-oriented purposes, and
- satisfying source and text-quality requirements

were chosen. It was also desirable to identify multiple sources for each medium so that the corpus would contain sufficient variation in language. Roughly equivalent proportions of words for each medium and for each genre factor were desired.

Collecting roughly equivalent word counts based on media factors proved less difficult than collecting equivalent proportions based on genre factors, because type of media is a mutually exclusive criteria category—e.g., either a text comes from a chat medium or it does not. So texts could be conclusively classified by type of medium, and obtaining roughly equal proportions of words for each of the five media was feasible.

To balance based on genre factors, sources were first classified by topic and purpose, and then as-equivalent-as-possible word counts were collected for each topic and purpose. Classifications were made by consulting documentation about the source. However, genre factor classification was less precise than classifying sources by medium because most documentation did not explicitly state topics and purposes; and in a few cases, there was no documentation at all for a particular source or its components (e.g., for a forum within a particular forum-providing source). So if documentation was nonspecific, anything known about participants or setting was also used to inform classification decisions, but decisions were, at best, inferences where the “most likely” dominant topic and purpose were selected. (See Appendix C: Support for Topic and Purpose Classifications for the evidence used to classify sources.)

For example, the University of North Carolina (UNC) University Libraries has no documentation about UNC Ask a Librarian IM (a selected source), and a review of the IM software’s website—LibraryH3lp—revealed no discussion of topic or of types of questions asked by library patrons. It did, however, provide some clues about purpose by explaining

that the service is to be used by libraries as a means of providing virtual reference services. Thus, it could be inferred from this minimal information about purpose, and knowledge about the participants (librarians, students, faculty, staff) and the setting (a university library), that conversation would be largely intended for serious purposes (i.e., related to work and school needs) and would most likely focus on academic and research-oriented topics related to library and information use (e.g., book or article finding assistance, citation practices, information needed to complete a homework assignment or to write a research article, etc.)—i.e., the Other topic.

Sometimes documentation alluded to both purposes—Serious and Recreational/Leisure-oriented purposes; and sometimes both documentation and information known about a source (e.g., participants, setting) did not provide strong evidence either way. However because the source satisfied other requirements, it was still deemed valuable. For these cases, two new purpose categories were needed: Mixed and Ambiguous. They are defined as follows:

- Mixed - The source's documentation or other information known about the source (e.g., types of participants, setting) does not suggest that is intended or used solely for one purpose over the other. Instead it appears to be used for **both** Serious and Recreational/Leisure-oriented purposes.
- Ambiguous - The source's documentation or other information known about the source (e.g., types of participants, setting) does not provide clear or strong evidence for either Serious or Recreational/Leisure-oriented purposes—i.e., purpose is unclear.²⁷

²⁷ In spite of the Ambiguous classification, Ambiguous sources were selected because they satisfied other requirements, such as media factors and source and text-quality requirements, and texts were easily obtained.

NCKnows IM (a selected source) is an example of the Mixed purpose. The website (<http://ncknows.org/aboutnc.html>) describes the service as follows:

NCKnows is a service that allows North Carolina residents to get help from librarians and use their library resources remotely through a computer. By "chatting" online with a librarian, you can get the most from your library, including access to articles, audiobooks and more from NC LIVE. Whatever you need, NCKnows will be able to get you started. It's free, helpful and easy. We've helped thousands of NC patrons over the years, including k12 students, business information seekers, college students, people looking for good books and many many more.

Although one of the participants in any IM conversation will be a librarian fulfilling his/her job duties (thus alluding to a more serious purpose), the description does mention that one of the reasons people may use the service is to find a good book. Whether the good book is used for serious or leisure activities is not clear, but the possibility of using it for recreation is high. Library patrons may use the service for both serious activities and recreational activities.

The Google Group for beekeepers is an example of the Ambiguous purpose classification. It was selected because it provided public access to emails, which was less difficult than collecting individual personal emails. Also, it met source and text-quality requirements by demonstrating sufficient participation, and it offered a non-gaming, non-technology Other topic. The only documentation was a single line explaining that the email list is a forum to discuss beekeeping and bees in a particular county in a west-coast state. Information about participants was not known, such as whether the beekeeping participants were using the email list for recreational purposes (i.e., beekeeping as a hobby) or serious purposes (i.e., beekeeping as honey-making and/or beeswax-product business). Thus, no

clear or strong evidence was available to confidently classify this email list as Serious, Recreational/Leisure-oriented, or Mixed.

Sinclair (2005a) explains that some criteria may be too ambiguous to draw fully reliable conclusions. Even with topic as a criterion—which appeared to be more conclusive than purpose and thus took greater precedence than purpose when sampling—one cannot say that a forum or chat room always discusses only one topic. Shifts in topics naturally occur in any conversation. This is particularly true of chat where multiple topics may be discussed at any given time, and thus cause the conversation to resemble “the randomness of the subject-matter in face-to-face conversation” (Crystal, 2006, p. 151). So although a source may be described in a way that suggests one particular topic, the reality may not bear out this designation. To continue the UNC IM example, one would expect few, if any, conversations about gaming (e.g., how to complete a quest, strategy for raiding a dungeon, etc.) or technology (e.g., how much memory a computer should have, which programming expressions to use to build specific functionality into a webform, which cloud storage services provide the best data security, etc.) that might be better suited to a help desk, but that does not preclude the possibility of these types of discussions. This is why the “most likely” qualification is needed. The most likely topic for UNC IM is the Other topic.

Another complication with genre factors that affects the ability to obtain equivalent word counts is that topics and purposes may not co-occur in equal proportions within a particular source or among sources. For example, all the IM sources selected for this study are library virtual reference services, which resulted in a somewhat homogeneous set of texts—i.e., virtual reference IM tends to focus on Serious or Mixed purposes and Other topics. However, feasibility concerns dictated the selection of IM sources. IM texts were

readily available in bulk from helpful librarians administering virtual reference services. Collecting personal IM messages from friends and other contacts or through general calls for messages may have resulted in lengthy collection times and potentially insufficient quantities of messages. It would also have required obtaining permissions from each participating interlocutor, which—if sufficient numbers of messages were obtained—could warrant obtaining permissions from hundreds or thousands of individuals. Library administrators controlling access to virtual reference IM could quickly pull thousands of conversations from an archive and scrub the messages for personally identifying information, thereby negating the need for individual permissions.

This lack of equivalent co-occurrence and the speculative nature of topic and purpose classifications means that proportions based on topic and purpose are not entirely dependable in terms of equivalence. While it was hoped that genre factor proportions would be roughly equal, any seemingly equal proportions cannot be said to be so with complete certainty because of these issues. Proportional sampling is “admittedly more of a theoretical ideal” (Gries, 2009, p. 1232). However, because conclusive topic and purpose classifications and reliably equivalent proportions were not possible with a high degree of certainty, genre factors assumed a secondary role to media factors in informing corpus creation decisions.

As was mentioned earlier, 150,000 word tokens were desired with roughly equivalent proportions—30,000 words—for each of the five media: forums, email, SMS, IM, and chat. This was achievable except in the case of SMS. Only one predominantly English source—Dr. Susana Sotillo’s SMS corpus—was available at the time of data collection. The NUS SMS Corpus (<http://wing.comp.nus.edu.sg/SMSCorpus/>) was available for use, but much of it includes messages in non-English languages. Therefore, the corpus does not contain a full

150,000 word tokens because Dr. Sotillo's corpus, at the time of collection, was 10,918 words, but other media contain the desired 30,000, and it was possible to draw from multiple sources for each of these other four media. The corpus thus contains 136,529 word tokens, and spans 12 sources (17 if each email list is considered a separate source).

Corpus building is “an inexact science, and no-one knows what an ideal corpus would be like” (Sinclair, 2005b, p. 81). “Given constraints on time, finances, and availability of texts, compromises often have to be made. Every corpus will have limitations, but a well-designed corpus will still be useful for investigating a variety of linguistic issues” (Biber et al., 1998, p. 250). Although this corpus has its limitations, as all corpora do, it is hoped that it is indeed useful for investigating language variation in online social media.

The remainder of this section provides a detailed outline of why sources were selected and how texts were collected. It discusses the constraints on data collection and any resulting limitations. Then the section concludes with details about how the corpus was cleaned in preparation for analysis.

Forums

Four forum sources were selected: Teenspot forums, Yahoo forums, EverQuest forums, and World of Warcraft (WoW) forums. All are publicly available for viewing on the web, and thus do not pose any barriers to collection. Roughly half of the desired 30,000 for forums—15,000—was desired for Other topics, and half for Gaming (7,500) and Technology (7,500) topics combined.

Teenspot claims to be the “premiere entertainment and community website for teenagers.”²⁸ Teenspot offers closed chat rooms, open discussion boards, news, polls, personal profiles, and more. Most Teenspot userids are not real names, and most profile pictures do not portray the user. For example, one user has a profile picture of Samuel Jackson from the movie *Pulp Fiction*. Teenspotters have the option to insert graphic emoticons (e.g., 😊) and other images into their forum messages, and Teenspot forums provide quoting functionality. There are Teenspot forums for newcomers to the website, general discussion, and topics such as sports, gaming, TV and movies, and celebrities. Teenspot was selected because it provides forums on a number of different topics (including gaming and technology), and because news stories and research studies highlight teen involvement in the creation of new linguistic forms in online settings. For example, the ABC Good Morning America website features online articles about “the secret language of teens” (Murphy & Allen, 2007) and “how to decode slang your teen uses online” (Murphy, 2010). Five Teenspot forums were selected: General, School, College, Technology, and Gaming. General, School, and College were classified as the Other topic and the Mixed purpose. The School and College forums, being similar in nature, were treated as one forum, with half the desired word counts being taken from each. The Technology and Gaming forums were classified as the Technology and Gaming topics respectively. Technology was classified as Ambiguous due to Teenspot’s ambiguous description of the forum. Gaming was classified as the Recreational/Leisure-oriented purpose.

Like Teenspot, Yahoo offers a variety of online services, which include search, news, personal profiles, email, topical webpages, and opportunities for online communication, such

²⁸ See <http://www.teenspot.com/about/>

as their forums. Yahoo forums were selected because Yahoo is a well-known online entity with household-name status, and offers an extensive listing of forums focused on a variety of topics, including business and finance, computer and Internet, family and home, health and wellness, science, etc. Many Yahoo userids do contain personally identifying information—i.e., many are email addresses—but non-identifying userids are possible. Use of profile pictures is variable. Yahoo forums do not support the inclusion of quoted pieces of previous messages in one’s reply, but users can reply to a specific post among the many that a thread may contain²⁹. Four forums were selected: Schools & Education, Family & Home, Computers & Internet, and Games. As with Teenspot, these Yahoo forums offer a mix of purposes and span all three topics. Schools & Education and Family & Home were classified with the Other topic. Computers & Internet and Games were classified with the Technology and Gaming topics respectively. The Family & Home and Games forums were classified with the Recreational/Leisure-oriented purpose, Schools & Education with the Serious purpose, and Computers & Internet with the Ambiguous purpose because no documentation was provided about the forum and no other information known about the forum definitively suggested one of the other purposes.

EverQuest and World of Warcraft are two gaming sources that provide forums to their players. EverQuest, developed by Sony Entertainment, is a well-established—13 years old at the time of this writing—MMOG.³⁰ World of Warcraft, developed by Blizzard, is the most popular massively multiplayer online game, boasting 12 million subscribers (Blizzard

²⁹ Many forums permit users to reply to the entire thread rather than selecting one post and replying only to that. In these forums, users can nevertheless respond to a particular message by including a quoted section of a message or an entire quoted message in their own posts. With Yahoo forums, replying to a specific post is the only way to respond to an individual message.

³⁰ See <https://www.everquest.com/faq>

Entertainment, 2010). It holds the Guinness World Record for the most popular MMOG (Samuel, 2011). Because both these games have solid footing in the gaming world, they were selected to strengthen the Gaming topic category. Additionally each has one or more technology-oriented forums, which add to the collection of technology-related conversations.

EverQuest offers players an opportunity to complete quests in a fictional, virtual world, and to socialize and collaborate with other players. The listing of forums at the time of data collection in March of 2011 has since changed, but its current offerings are not all that dissimilar. It offers technical support forums, forums focusing on helping newcomers to the game, forums for specific character classes,³¹ and more. Most participants use their character names—usually not personally identifying—as their userids, and profile pictures are rarely of the actual player. EverQuest forums support quoting of previous messages. Four forums were selected: Gameplay Content, The Newbie Zone, The Veteran’s Lounge, and Gameplay Mechanics, which was a technical support forum and allowed the collection of conversation focused on Gaming-technology topics.

Like EverQuest, Blizzard has changed the listing of WoW forums since data collection in March of 2011, but current offerings are similar in nature. As in EverQuest, userids tend to be character names. Profile pictures tend to be cartoons of a player’s character and quoting is supported. Five forums were selected: General, New Player Help and Guides, Quests, Technical Support, and Mac Technical Support. Like the School and College forums from Teenspot, Technical Support and Mac Technical Support were treated as one forum because of their similarity. They provide conversation on Gaming-technology topics.

The process outlined below is a general overview of how texts were selected from these forum sources. Exceptions, based on a particular source’s structure or focus, are noted

³¹ A character class is a role that one’s character plays. Examples include mages, priests, warriors, etc.

both in the steps below and the subsequent description of the collection process. Four forums per source were desired to provide a sufficient range of topics and purposes, and only texts that contained posts from the 10 years prior to collection were obtained.

1. Forums and/or topic categories that offered discussion of Gaming and Technology topics, and that satisfied the source and text-quality requirements, were selected first. (Teenspot and Yahoo offered only one forum or topic category for each. All WoW and EverQuest forums focus on gaming topics. EverQuest had one technical support forum. WoW had two—each pertaining to a different operating system, so both were selected, and half the needed word counts to represent Technology in WoW forums were obtained from each. These three EverQuest and WoW forums were viewed as providing conversation on Gaming-technology topics.) If the forum offered some type of “general” forum, this was also selected as either a representative of the Other topic as in the case of Teenspot, or as a representative of greater breadth of topic as in the case of WoW forums. This step resulted in three forums from Teenspot (General, Technology, Gaming), two forums from Yahoo (Computers & Internet, Games), one forum from EverQuest (Gameplay Mechanics), and two from WoW that were to be treated as one (Technical Support and Mac Technical Support).
2. If more forums were needed to complete the list of four desired (i.e., usually to fill out the Other topic category because there was no “general” forum available, or in the case of WoW and EverQuest, to select additional non-technology forums), the remaining forums and topic categories were examined to ensure they satisfied source and text-quality requirements. Those that did not satisfy these requirements were excluded from further consideration. EverQuest and WoW forums intended for

Blizzard employees to broadcast news were also excluded. These forums did not permit responses from members of the gaming community, and so did not satisfy the definition of conversation.

3. Random numbers were used to select the remaining forums. However, if only one forum was needed, then the first available forum that satisfied all requirements was selected.
4. Once a forum was selected, threads with at least five posts were selected in the order they appeared on the page, which was usually chronological with threads that had the most recent replies to an original post listed first. Thread texts were collected until the desired word count was reached. To improve the chances of collecting as much variation in language as possible, usually no more than 15 posts per thread were collected before moving on to another thread in the forum.

Teenspot, EverQuest, and WoW forums structure their forums similarly. For example, at the time of data collection, Teenspot forums included 30 forums listed on the Boards webpage. Upon clicking on one of these forums, users are presented with a list of threads. Clicking on a thread takes the user to the posts on that thread's topic. So three easy clicks to conversation, in a simple, hierarchical, linear, organizational scheme (i.e., threads were not listed under multiple topic headings—there was only one way to reach a particular conversation). But Yahoo forums are nested, with a network-style organization. For example, the initial listing on the Yahoo! Message Boards page contains a listing of topic categories³², not links to actual forums. Each topic category may link to more topic categories and/or actual forums. Users must drill down through several pages before reaching conversation,

³² A *topic category* is defined as a subject heading that links to a list of forums or other topic categories classified under the initial topic category.

and forums may be listed under multiple topic categories (i.e., so there were multiple ways to reach a conversation). This variation in structure affected the sampling process outlined above because it required digging into multiple levels of links to get to forums that satisfied source and text-quality requirements.

For Yahoo, step 1 above was roughly the same. Top-level topic categories that matched Gaming and Technology topics were selected. Because there was no “general” topic category on the main forums page, steps 2 and 3 were applied to the selection of the top-level topic categories to select the remaining two Other forums. Then, when drilling into a topic category, random numbers were used to select from subsequent lists of topic categories and/or forums until reaching a forum that satisfied source and text-quality requirements. For example, most forums within the Business & Finance topic category did not contain real conversation; most were lists of job ads. So these forums were not selected.

Participation can be quite sparse in Yahoo forums, so satisfying the objective of sufficient participation had great influence on the final selection of texts. The no-more-than-15-posts guideline discussed in step 4 above could not be applied to most Yahoo threads because few threads had even five posts. Oftentimes, the rare threads with as many as 25 posts were selected to complete the desired word counts for Yahoo.

Table 4³³ below lists word counts and average words per conversation and message for all forums sampled.

³³ All word counts listed in the Corpus Creation section were calculated with TextWrangler’s (<http://www.barebones.com/products/textwrangler/>) word count function.

Table 4: Word counts for the forums section of the corpus.

Forum	Proposed N Words to Collect	N Conversations (Threads) Collected	N Individual Messages Collected	N Words Collected	Average Words Per Conversation	Average Words Per Message
Teenspot						
General	3,750	12	166	3,823	319	23
1/2 School and 1/2 College	3,750	9	130	4,049	450	31
Technology	1,875	6	65	2,031	339	31
Gaming	1,875	6	65	2,000	333	31
Totals	11,250	33	426	11,903	361	28
Yahoo						
Schools & Education	3,750	3	56	4,517	1,506	81
Family & Home	3,750	5	52	3,374	675	65
Computers & Internet	1,875	2	29	2,156	1,078	74
Games	1,875	6	82	1,920	320	23
Totals	11,250	16	219	11,967	748	55
EverQuest						
Gameplay Mechanics	1,875	3	45	1,915	638	43
Gameplay Content	625	1	15	1,260	1,260	84
The Newbie Zone	625	1	15	458	458	31
The Veteran's Lounge	625	1	12	621	621	52
Totals	3,750	6	87	4,254	709	49
World of Warcraft						
1/2 Technical Support and 1/2 Mac Technical Support	1,875	5	45	2,107	421	47
General Discussion	625	2	21	708	354	34
New Player Help and Guides	625	1	15	626	626	42
Quests	625	2	22	1,072	536	49
Totals	3,750	10	103	4,513	451	44
All forums						
Totals	30,000	65	835	32,637	502	39

Email lists

The barriers to collection of IM conversations, discussed at the beginning of the Corpus Creation chapter, are also barriers for the collection of individual email. To obtain enough variation in language, it would be desirable to obtain messages from a wide variety of users, and both senders' and receivers' permissions would be required. Depending on how

many interlocutors a corpus creator wishes to have represented in the corpus, locating enough willing participants and obtaining the necessary permissions may result in a lengthy collection period. Although collecting from friends and family and their associates might reduce collection times and make permission acquisition easier, it may yield a rather homogeneous set of language—birds of a feather flock together, so the saying goes. Furthermore, messages may be edited or filtered in some way before participants submit them to the researcher, and this may result in less than pure language samples.

Publicly available email lists, however, offer a nice alternative to individual email by eliminating many of these issues. Google Groups offers users the opportunity to form discussion groups and communicate with these groups through email. Many Google Groups are publicly available for viewing; they do not require a userid and password to read a group's messages. Additionally, at the time of collection, Google classified groups according to topic, number of messages per month, geographic location, language, number of members, etc. The classification of email lists by topic permitted the selection of email lists that fit the desired Gaming, Technology, and Other topics, and Google Groups' extensive collection of email lists—both professionally and recreationally-oriented—permitted the collection of lists spanning both Serious and Recreational/Leisure-oriented purposes. Like Yahoo forums, Google Groups are organized network-style, with groups appearing under multiple topic categories. Six Google Groups were selected for this corpus: a multiple sclerosis support group, a group of transcriptionists, fans of an opera singer, beekeepers, computing experts, and players of a specific multiplayer game. The computing experts and the game lists fulfilled the need for Technology and Gaming topics respectively, and the other lists satisfied the need for conversation about Other topics. The multiple sclerosis, transcriptionists', and

computing lists fulfilled the need for Serious conversation. The email list for fans of an opera singer was classified as having the Recreational/Leisure-oriented purpose, and the beekeeper list was classified with the Ambiguous purpose. Roughly half of the 30,000 desired for email lists—15,000—was desired for the email lists that focused on Other topics, and half was desired for the combination of Gaming (7,500) and Technology topics (7,500).

The process for selecting these email lists was similar to the selection of Yahoo forums. The topic categories listed on the main Google Groups' webpage that appeared to offer discussion of Gaming and Technology topics were selected. Then, the resulting list of groups was filtered so that only those with 10-99 messages per month were listed. This new list of groups was then reviewed, and instead of using random numbers, the first email list that satisfied criteria was selected. This was because initial attempts at using random numbers—of which there were several—proved to be clumsy and slowed the process considerably. Using random numbers resulted in repeated selection of email lists that were either closed or did not completely satisfy genre or source and text-quality requirements. For example, under the Computers category, an email list for recruiters was listed; however, computers and other facets of technology were not actually discussed. Instead the email list seemed to be primarily for posting open jobs in the information technology field. So the email list acted more as a job board than an arena for technology discussion, and was therefore not selected. After numerous failed attempts to select viable email lists, it became more expeditious to go through the list on a case-by-case basis and select the first email list that satisfied requirements.

To select email lists that fulfilled the need for discussion of Other topics, each of the other top-level topic categories was evaluated in the order they appeared on the main Google

Groups webpage in a fashion similar to the process for selecting gaming and technology forums. A category was clicked. The resulting list was further filtered by 10-99 messages per month. Then, as with Gaming and Technology email lists, the first email list that satisfied criteria was selected. If more than three pages of email lists were examined and none were found that satisfied requirements, then another topic category was selected from the main Google Groups webpage and the process was repeated.

Similar to forum thread selection, once a group was selected, conversations with at least five messages were selected—i.e., conversations were selected where one person sent a message to the list and at least four responses were emailed back to the list. Conversations were also selected in the order they appeared on the page, with the most recent messages listed first. Conversations were collected until the desired number of words was reached.

Table 5 below lists word counts and average words per conversation and message for all lists sampled.

Table 5: Word counts for the email lists section of the corpus.

Email List	Proposed N Words to Collect	N Conversa- -tions Collected	N Individual Messages Collected	N Words Collected	Average Words Per Conver- -sation	Average Words Per Message
A multiple schlerosis support group	3,750	7	60	3,747	535	62
A group of transcriptionists	3,750	7	68	4,308	615	63
Fans of an opera singer	3,750	4	26	3,984	996	153
Beekeepers	3,750	10	81	4,636	464	57
Computing experts	7,500	6	71	7,508	1,251	106
Multiplayer game players	7,500	12	106	7,676	640	72
Totals	30,000	46	412	31,859	693	77

SMS

The process of obtaining text messages presents the same barriers as with IM and email. So it was desirable to find a pre-existing corpus of SMS messages. At the time of data collection, there was only one SMS corpus freely available to researchers: the NUS SMS corpus. A rather large corpus, NUS would have easily provided the full 30,000 words desired for the SMS medium, but a large number of messages contained non-English utterances and code-switching between English and other languages spoken in Singapore. Because the researcher does not read non-English languages fluently and was unfamiliar with Singaporean languages, the NUS SMS corpus was not a viable source.

Efforts turned toward asking members of the corpora-list (<http://www.hit.uib.no/corpora/>) if they knew of other predominantly English SMS corpora. Only one member knew of another corpus—her own that she was in the process of collecting and cleaning. Dr. Susana Sotillo in the Department of Linguistics at Montclair University graciously agreed to share her SMS corpus with the researcher. Unfortunately, because it was in the early stages of development and was still being de-identified by Dr. Sotillo, a full 30,000 words was not possible. Dr. Sotillo was only able to share 10,918 words of her corpus. However, the corpus is predominantly in English and offers utterances from 59 individuals, including teenagers and adults, using SMS for both personal and business reasons (Sotillo, 2010). Thus, the Sotillo corpus satisfied the need for English conversation spanning a variety of Other topics used for both serious and recreational purposes (i.e., Mixed classification). The corpus in its entirety was appropriated.

Table 6 below lists word counts and average words per message for the Sotillo corpus.

Table 6: Word counts for the SMS section of the corpus.

SMS	Proposed N Words to Collect	N Individual Messages (Lines) Collected	N Words Collected	Average Words Per Message (Lines)
Sotillo Corpus	30,000	1,391	10,918	8

IM

As has been said, IM presents certain barriers to data collection, largely centered on permission issues, risk of homogeneity if collecting from known associates, and lengthy collection times. Virtual reference IM solves many of these problems. Libraries already have policies in place that permit the use of transcripts in research, as long as identities are scrubbed. Thousands of messages may be obtained in bulk; and although using only virtual reference sources for IM results in a somewhat homogeneous set of topics (Other) and purposes (Serious and Mixed—though leaning more toward Serious than Recreational/Leisure-oriented), the interlocutors themselves are not homogeneous and so language may be more varied than with birds-of-a-feather situations such as sampling from friends and family.

Three virtual reference IM sources were selected: UNC Ask a Librarian, NCKnows, and L-net. Through a self-made contact and referrals from a professor, the researcher was able to establish contact with the librarians administering these services and obtain their permission to use the data. Because of their similarity, roughly 10,000 words from each of these sources were desired to complete the 30,000 desired for IM.

UNC Ask a Librarian is a service offered by the University Libraries at the University of North Carolina at Chapel Hill. University students, faculty, and staff may use the service to obtain help with research and information-seeking questions. It is staffed by professional

and student librarians who specialize in reference. The service can be accessed through the University Libraries webpage or other IM programs like AIM (AOL Instant Messenger).

NCKnows is North Carolina's statewide virtual reference service, thus any North Carolina resident has access to this service, including "K-12 students, business information seekers, college students, [and] people looking for good books."³⁴ It is staffed by professional librarians working in academic and public libraries across the state. As with UNC, NCKnows provides help with information seeking questions.

L-net is Oregon's statewide virtual reference service, thus any Oregon resident can use the service. It is staffed by volunteers and professional librarians working in academic and public libraries, both within and outside of Oregon. L-net librarians help with homework, research, book finding, verifying citations, and more.

Sampling was based on level of activity per month. UNC IM was collected first, and NCKnows and L-net sampling was modeled on it. Texts were collected from UNC in March 2009. Because the request for texts fell outside of the librarian's regular duties, it was important to make the parameters of the request as simple as possible so as not to apply additional strain on an already busy work schedule. Thus what was most convenient for UNC was to provide conversations from 5/15/07 – 5/15/08—the year of IMs logged before switching to a new IM backbone at the end of 2008, which made retrieval of archived messages more difficult. Based on the average word counts of a few personal IMs and the desire not to overwhelm the UNC librarian who would need to scrub the messages of identifying information (e.g., userids, phone numbers, email addresses), it was decided that 300 conversations would most likely yield a sufficient number of words for this source. The librarian presented the researcher with a list of all conversations for that year. This list

³⁴ See <http://ncknows.org/aboutnc.html>

consisted of the conversation's ID and date. It did not contain conversation content. There were 2,541 conversations that year, and 300 represented 12% of them.

Then the researcher used random numbers to sample conversations proportionate to the level of activity per month. For example, May 2007 had a total of 24 messages (.0094 of the 2,541 conversations in that year). So three conversation IDs (.0094 of the desired 300 conversations) were selected at random. Once all 300 conversation IDs were selected, the researcher presented the list of them to the librarian who then had the conversations scrubbed before presenting them to the researcher.

Ultimately only a subset of these 300 conversations was needed to fulfill the desired word count of 10,000. The proportions used earlier in the process to obtain the original 300 were also used to determine the proportion of words needed from each month to fill the 10,000. For example, 0.0094 of 10,000 equals, roughly, 94 words; so for May 2007, one or more May conversations were selected based on how closely their word counts matched the desired 94 words. This was done for each month in that year. Conversations were not truncated to achieve a perfect 10,000 words; each conversation in its entirety was used. Table 7 below shows the number of conversations that were available for sampling during the period of 5/15/07 through 5/15/08. Then the proportion that those conversations assumed out of the total conversations available for collection (2,541) is displayed by month. The number of conversations that were obtained—based on those monthly proportions—is listed, as well as the number of words desired based on monthly proportions. Figures in Table 7 have been rounded to the fourth decimal point.

Table 7: UNC Ask a Librarian sampling statistics.

	Number of Conversations Available for Sampling	Proportion of Conversations out of Total Conversations Available (e.g., 24/2,541)	Number of Conversations Proportionate to the Level of Monthly Activity (e.g., 0.0094 x 300)	Number of Words to Sample to Complete the Desired 10,000 Words (e.g., 10,000 X 0.0094)
May-07	24	0.0094	2.8335	94.4510
Jun-07	50	0.0197	5.9032	196.7729
Jul-07	55	0.0216	6.4935	216.4502
Aug-07	93	0.0366	10.9799	365.9976
Sep-07	300	0.1181	35.4191	1180.6375
Oct-07	278	0.1094	32.8217	1094.0575
Nov-07	348	0.1370	41.0862	1369.5396
Dec-07	172	0.0677	20.3070	676.8989
Jan-08	242	0.0952	28.5714	952.3810
Feb-08	280	0.1102	33.0579	1101.9284
Mar-08	221	0.0870	26.0921	869.7363
Apr-08	397	0.1562	46.8713	1562.3770
May-08	81	0.0319	9.5632	318.7721
Total	2,541	1.0000	300	10,000

NCKnows conversations were obtained from Dr. Jeffrey Pomerantz, an associate professor at UNC Chapel Hill’s School of Information and Library Science, who had conducted an earlier study of the NCKnows service. The stipulations of Dr. Pomerantz’s data use agreement with NCKnows allowed him to share the data with other researchers. The data Dr. Pomerantz collected included two years of IM conversations (2004-2005), but for this study, only the most recent year’s (2005) conversations were used.

Because 12% of the total messages available from UNC were collected, 12% (1,378) of the 2005 NCKnows messages (11,487) were also collected, and the rest of the NCKnows sampling process replicated the process for collecting UNC IM conversations. For example, in January 2005, there were a total of 1,065 conversations (.0927 of the 11,487 conversations

that year). So 127 conversations (.0927 of the desired 1,378 conversations) were selected at random. Once all 1,378 conversations were selected from the database of conversations given to the researcher by Dr. Pomerantz, a subset was selected to fulfill the desired word count of 10,000. So 927 words were selected from January 2005, for example. Conversations were not truncated. Table 8 shows the number of conversations that were available for sampling, the proportion that those conversations assumed out of the total conversations available, and the number that the researcher hoped to collect.

Table 8: NCKnows sampling statistics.

	Number of Conversations Available for Sampling	Proportion of Conversations out of Total Conversations Available (e.g., 1,065/11,487)	Number of Conversations Proportionate to the Level of Monthly Activity (e.g., 0.0927 x 11,487)	Number of Words to Sample to Complete the Desired 10,000 Words (e.g., 10,000 X 0.0927)
Jan-05	1065	0.0927	127.7592	927.1350
Feb-05	1248	0.1086	149.7122	1086.4455
Mar-05	1243	0.1082	149.1124	1082.0928
Apr-05	1166	0.1015	139.8753	1015.0605
May-05	785	0.0683	94.1699	683.3812
Jun-05	771	0.0671	92.4905	671.1935
Jul-05	659	0.0574	79.0548	573.6920
Aug-05	613	0.0534	73.5365	533.6467
Sep-05	1026	0.0893	123.0807	893.1836
Oct-05	1170	0.1019	140.3552	1018.5427
Nov-05	1028	0.0895	123.3206	894.9247
Dec-05	713	0.0621	85.5327	620.7017
Total	11,487	1.0000	1,378	10,000

L-net conversations were collected in February 2011. The process for selecting L-net conversation matched that of UNC and NCKnows. However, because L-net's volume of

conversations is substantially higher than UNC or NCKnows, only a six month sampling frame was requested: 8/15/10 – 2/15/11. Within that six months there were a total of 13,914 conversations. Instead of taking a full 12% of those, half was requested (6% or 835) to match the halved sampling frame. Conversations were not truncated. Table 9 below shows the number of conversations that were available for sampling, the proportion that those conversations assumed out of the total conversations available, and the number that the researcher hoped to collect.

Table 9: L-Net sampling statistics.

	Number of Conversations Available for Sampling	Proportion of Conversations out of Total Conversations Available (e.g., 648/13,914)	Number of Conversations Proportionate to the Level of Monthly Activity (e.g., 0.0466 x 835)	Number of Words to Sample to Complete the Desired 10,000 Words (e.g., 10,000 X 0.0466)
Aug-10	648	0.0466	38.8800	465.7180
Sep-10	1617	0.1162	97.0200	1162.1389
Oct-10	2812	0.2021	168.7200	2020.9861
Nov-10	2745	0.1973	164.7000	1972.8331
Dec-10	1587	0.1141	95.2200	1140.5778
Jan-11	2994	0.2152	179.6400	2151.7896
Feb-11	1511	0.1086	90.6600	1085.9566
Total	13,914	1.0000	835	10,000

Table 10 below lists collected word counts and average words per conversation and line for all IM sources sampled.

Table 10: Word counts for the IM section of the corpus.

IM Conversation	Proposed N Words to Collect	N Conversations Collected	N Individual Messages (Lines) Collected	N Words Collected	Average Words Per Conversation	Average Words Per Message (Line)
UNC						
May-07	94	1	12	135	135	11
Jun-07	197	1	21	211	211	10
Jul-07	216	1	41	260	260	6
Aug-07	366	1	31	402	402	13
Sep-07	1,181	2	95	1,254	627	13
Oct-07	1,094	5	98	1,069	214	11
Nov-07	1,370	6	157	1,371	229	9
Dec-07	677	2	102	695	348	7
Jan-08	952	5	121	958	192	8
Feb-08	1,102	5	153	1,133	227	7
Mar-08	870	4	89	894	224	10
Apr-08	1,562	6	187	1,575	263	8
May-08	319	2	39	327	164	8
Totals	10,000	41	1,146	10,284	251	9
NCKnows						
Jan-05	927	5	75	972	194	13
Feb-05	1,086	6	72	1,157	193	16
Mar-05	1,082	6	81	1,098	183	14
Apr-05	1,015	4	99	1,069	267	11
May-05	683	3	62	767	256	12
Jun-05	671	3	41	703	234	17
Jul-05	574	2	33	575	288	17
Aug-05	534	3	62	568	189	9
Sep-05	893	4	86	1,001	250	12
Oct-05	1,019	6	79	1,083	181	14
Nov-05	895	4	101	969	242	10
Dec-05	621	3	34	628	209	18
Totals	10,000	49	825	10,590	216	13
L-net						
Aug-10	466	3	45	469	156	10
Sep-10	1,141	7	118	1,143	163	10
Oct-10	1,973	13	212	1,974	152	9
Nov-10	2,021	12	320	2,027	169	6
Dec-10	1,162	5	93	1,201	240	13
Jan-10	1,086	6	111	1,096	183	10
Feb-10	2,152	15	202	2,170	145	11
	10,001	61	1,101	10,080	165	9
All IM						
Totals	30,001	151	3,072	30,954	205	10

Chat

Three chat sources were selected for inclusion in the corpus: AOL chat, WoW chat, and the NPS chat corpus. AOL chat and NPS chat provided conversation about Other topics; World of Warcraft chat was focused on Gaming topics. All three were for Recreational/Leisure-oriented purposes. Roughly half of the desired 30,000 for chat—15,000—was desired for Other topics (AOL and NPS) and half for Gaming topics (WoW). These three chat sources posed few barriers to collection.

AOL chat was offered by America Online (AOL), an online entity akin to Yahoo in many respects. AOL offers search functionality, news, topical webpages, personal profiles, and opportunities for communication, such as their chat rooms. AOL chat provides conversation on Other topics for Recreational/Leisure-oriented purposes. AOL chat has since been discontinued while AOL revamps the chat service, but at the time of data collection, AOL chat rooms could be viewed and chat could be copied without logging in. Furthermore, AOL quickly and graciously granted permission to use their chat for this research. AOL userids may have contained personally identifying information but most tended to disguise the interlocutor's identity, and as with all other parts of the corpus, the researcher agreed to de-identify the text before sharing the corpus with others.

AOL chat was collected in May and June of 2009 during each day of the week. Individual chat rooms were selected based on strong participation, and low incidence of sex bots and solicitation. Fifteen rooms were thus sampled, many repeatedly, throughout these two months. Four-hour increments were obtained during three segments of each day: morning (6am-12pm), afternoon (12pm-6pm), and evening (6pm-12am).³⁵ Additionally, if

³⁵ Increments were not taken from a nighttime segment of 12am-6pm because this is when the researcher slept.

the researcher needed to go to an appointment, chat was not collected during that segment. If a conversation in a particular chat room died out before the four-hour increment was over, another chat room was immediately sampled. Chat was copied from the chat window in the web browser, and pasted into a text file. A high volume of chat was ultimately collected, but like the IM sources, only a subset was needed to fulfill the desired word count of 7,500. Thus, the first 500 words from the most recent chat files for each of the rooms sampled became a part of the final corpus analyzed for the dissertation. Utterances were not truncated at the 500-word mark. So if the first 500 words ended in the middle of an interlocutor's message, the entire message was included.

Ultimately, multiple game-based chat sources were desired to flesh out the Gaming chat portion of the corpus, rather than just World of Warcraft chat alone. However, at the time of data collection, the researcher had established skill and participant legitimacy³⁶ with only World of Warcraft. The intimate and in-depth knowledge of World of Warcraft gained from years of playing the game made the collection of rich conversation possible. For example, the researcher played characters at all experience levels—characters that were in the early stages of the game all the way up to characters that had reached the level cap; so she was able to encounter and capture conversation with a wider variety of other WoW players about a wider variety of game tasks and topics.

Blind chat logging through the use of a bot or an inactive character (parked in a heavily populated area) that fails to respond to communication may result in superficial, one-sided communication. Also, inactivity may arouse suspicion and distrust. For instance, inactive players in player versus player battlegrounds are frequently suspected of being bots,

³⁶ See Lave, J. & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

and the other members of the team may choose to remove the player from the team and the match. According to Cherny (1999, p. 303), “lurking without participating is noticeable and ultimately difficult.” Furthermore, using bots to collect data is often prohibited and could result in an indefinite ban from the game. As the corpus grows in the future, the researcher will seek to broaden the scope of Gaming chat section to include chat from other games.

Access to WoW requires login with a userid and password. At the time of data collection, Blizzard provided instructions, on its public website, on how to log chat to a text file on the player’s computer using the */chatlog* command. In effect, this served as fair warning to all players that other players may log any conversation. The UNC Institutional Review Board found this to be sufficient evidence in favor of viewing WoW chat as public and therefore permitted the researcher to capture WoW chat. Additionally, as with all parts of the corpus, the researcher agreed to de-identify interlocutor userids in any publications even if the userid is not personally identifying.

The researcher used the game add-on Chatter (<http://www.wowace.com/addons/chatter/>) to log chat. It uses the */chatlog* command behind the scenes, automatically logging chat and thus eliminating the need to re-input the */chatlog* command every time one logs in. Players, including the researcher, often log in and out repeatedly during game play to switch characters they play or to adjust game settings. Forgetting to re-input the */chatlog* command at each login and miss out on valuable chat logging was a risk that Chatter eliminated.

By April 2011, the researcher had two WoW chat logs collected from two U.S. Servers: (a) from 12/1/08 through 4/5/09 on one server, and (b) 10/19/10 through 4/8/11 from log A’s server and a new server. Fifteen thousand word tokens were desired from WoW chat.

Half was taken from log A and half from log B. Because log A's chat was more than two years old, the last 7,500 words were taken. These came from March and April 2009. For log B, equivalent proportions of words from each month in that four-month time period were selected. Because October was essentially half a month of chat, the first 682 words were selected. From all remaining months except April, the first 1,364 words were taken. Because April included only a small number of words from one day—April 8th—it was not sampled. Utterances were not truncated at the 7,500-word mark; full utterances were retained.

Originally the researcher attempted to collect chat from Yahoo chat rooms; however, the requirement of sufficient participation was woefully unmet. So permission to use the NPS chat corpus—freely available to researchers—was obtained. NPS provides chat conversation about Other topics for recreational purposes. It contains 10,597 messages from a variety of chat rooms that are delineated by age (e.g., chat rooms for teens, 20-year-olds, 30-year-olds, etc.). Texts are part-of-speech and dialogue-act tagged. NPS corpus creators removed personally identifying information before making the corpus publicly available. Not all chat was needed to fulfill the desired word count of 7,500. There were 15 files in the NPS corpus. So, as with AOL chat, the first 500 words from each file in the NPS corpus was sampled and included for analysis in this study. As with AOL chat, utterances were not truncated at the 500 word mark. Instead the full utterance was retained.

Table 11 below lists word counts and average words per conversation and line for all chat sources sampled.

Table 11: Word counts for the chat section of the corpus.

Chat Room	Proposed N Words to Collect	N Individual Messages (Lines) Collected	N Words Collected	Average Words Per Message (Lines)
World of Warcraft				
12/08 - 4/09	7,500	1,017	7,502	7
10/10 - 4/11	7,500	1,021	7,536	7
Totals	15,000	2,038	15,038	7
AOL				
Room A	500	139	503	4
Room B	500	107	502	5
Room C	500	108	500	5
Room D	500	111	510	5
Room E	500	73	502	7
Room F	500	137	510	4
Room G	500	93	509	5
Room H	500	137	501	4
Room I	500	126	505	4
Room J	500	89	503	6
Room K	500	172	502	3
Room L	500	127	500	4
Room M	500	114	501	4
Room N	500	75	506	7
Room O	500	119	503	4
Totals	7,500	1,727	7,557	4
NPS				
Teens room A	500	75	500	7
Teens room B	500	124	505	4
Teens room C	500	109	500	5
20s room A	500	113	504	4
20s room B	500	116	500	4
20s room C	500	84	512	6
30s room	500	123	503	4
40s room A	500	89	506	6
40s room B	500	99	502	5
40s room C	500	78	501	6
40s room D	500	97	501	5
Adults room A	500	101	501	5
Adults room B	500	62	503	8
Adults room C	500	88	503	6
Adults room D	500	78	525	7
Totals	7,500	1,436	7,566	5
All chat				
Totals	30,000	5,201	30,161	6

Corpus cleaning

Once texts were collected, original versions of the texts were preserved as backups. Then copies were made and, using a series of regular expressions, these copies were cleaned of non-conversational text, such as date and time stamps, system messages, and userids. These copies were used to create word frequency lists used during analysis; and when terms needed to be examined in context to determine term meaning and features used, these copies were consulted.

The cleaning process varied for each source. For example, WoW chat required the removal of date and time stamps, chat channel information, userids, and game feedback. Below in Figure 1, is a hypothetical line of chat:

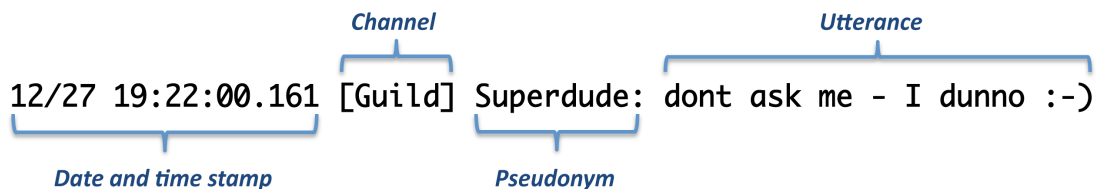


Figure 4: A hypothetical line of chat.

All but the “*dont ask me - I dunno :-)*” was removed. The WoW chat function automatically outputs text to the chat window in the form of feedback to the player about his/her activities. For example, the game will inform the player that s/he has been awarded game currency as a result of completing a quest (as in “*Received 7 Gold, 85 Silver*”). This game narrative was removed.


Similarly, the first line of most of the NPS chat files was a system message welcoming the interlocutor to the room and providing a link to chat policies. These were removed. Part-of-speech and dialogue act tags were a part of the XML code that housed NPS chat conversations. These tags and the XML code were removed. In forums and some

Google Groups messages, interlocutors may include special signatures. These signatures are usually configured while setting up a forum profile or through an email client's signature functionality. Then these signatures are automatically appended to each message the interlocutor sends. These were removed because they may contain—more often in the case of emails—personally identifying information, and they act more as identification badges one might be required to carry in the workplace, rather than conversation.

Other types of text were also removed, but replaced with certain codes. These were cases where it was believed that some indication of what was removed was necessary to preserve the integrity of the message, so that sufficient context was available to inform decisions about term classification by feature, during analysis. For example, when interlocutors used quoting functionality in forums, the quotes were substituted with the code *[Quoted_text]* so that the researcher would understand that some earlier comment was being referred to. This code, in particular, was also important for eliminating duplicates of existing comments that, if retained, would have led to erroneously inflated word counts. Other examples of substitution codes include *[Email_address]* for deleted email addresses, *[Phone_number]* for deleted phone numbers, and *[HTML_code]* for the few instances of HTML code that were deleted.

In contrast to using codes to stand in for deleted text, substitution codes were also used to stand in for conversational elements that could not be communicated via text but were worthy of retention and analysis. Specifically, the graphic emoticons in Teenspot were considered to be valuable conversational information, and substitution codes were used to retain that information even though retaining the graphics themselves was impractical.³⁷ The

³⁷ It was impractical to include the graphics in the messages because the Perl script used later in the process to create the word frequency list was not designed to process graphics, only text.

substitution codes were derived from the alt tags used to describe the emoticons. For example, the alt tag for this emoticon——was *EEK* and so the substitution code for this emoticon was *[EEK_emoticon]*.

Conclusion

The Institutional Review Board at the University of North Carolina at Chapel Hill (IRB 08-1886) approved the creation of this corpus. Data use agreements were required and obtained for:

- SMS messages (permission given by Dr. Susana Sotillo, Montclair University),
- UNC Ask a Librarian IM (permission given by Pam Sessoms, University of North Carolina),
- NCKnows IM (permission given by Dr. Jeffrey Pomerantz, University of North Carolina),
- L-net (permission given by Caleb Tucker-Raymond, L-net),
- AOL chat (permission given by AOL),
- NPS chat (permission given by Dr. Craig Martell, Naval Postgraduate School).

All other sources were considered to be publicly accessible by the UNC IRB, and so did not require express permission of the data owner/curator.

Once all texts were collected, the corpus consisted of 136,529 word tokens, almost 14,000 shy of the originally desired 150,000, due to the difficulty in obtaining another SMS source. With the exception of SMS, roughly equivalent proportions of words were collected for each of the five media; and healthy representation of the three main topics (gaming, technology, other) and two main purposes (serious, recreational) was achieved. Table 12

shows all 12 sources (or 17 if each email list is considered a separate source), the individual forums for each of the four forum sources, the dates messages were written, the most likely dominant topic and intended purpose, the number of words desired for each source (and forum), the total words collected for each source (and forum), and the total words collected for each medium.

Table 12: Corpus details.

Medium	Source	Dates Messages Written (or forum posts started)	Dominant Topic	Intended Purpose	Desired		Collected		
					Total Words to Collect	Total Words Per Forum	Total Words Per Source	Total Words Per Medium	
Forums	Teenspot (http://www.teenspot.com/boards/)								
	General	3/09, 1/10, 2/11, 3/11	Other	Mixed	3,750	3,823			
	1/2 School and 1/2 College	1/07, 3/08, 3/09, 4/09, 7/10, 10/10, 1/11, 2/11	Other	Mixed	3,750	4,049		11,903	
	Technology	1/07, 4/10, 9/10, 2/11	Technology	Ambiguous	1,875	2,031			
	Gaming	4/10, 2/11, 3/11	Gaming	Recreational	1,875	2,000			
	Yahoo (http://messages.yahoo.com)								
	Schools & Education	11/03, 3/04	Other	Serious	3,750	4,517			
	Family & Home	3/04, 4/05, 9/06, 3/09	Other	Recreational	3,750	3,374		11,967	
	Computers & Internet	12/03	Technology	Ambiguous	1,875	2,156			
	Games	8/01, 11/03, 2/04, 3/04, 7/04, 8/04	Gaming	Recreational	1,875	1,920			
	EverQuest (http://forums.station.sony.com/eq/forums/list.m)								32,637
	Gameplay Mechanics	3/11	Gaming Technology	Recreational	1,875	1,915			
	Gameplay Content	12/04	Gaming	Recreational	625	1,260		4,254	
	The Newbie Zone	12/10	Gaming	Recreational	625	458			
	The Veteran's Lounge	07/10	Gaming	Recreational	625	621			
	World of Warcraft (http://forums.worldofwarcraft.com/index.html?sid=1)								
1/2 Technical Support and 1/2 Mac Technical Support	11/10, 2/11, 3/11	Gaming Technology	Recreational	1,875	2,107				
General	3/11	Gaming	Recreational	625	708		4,513		
New Player Help and Guides	11/10	Gaming	Recreational	625	626				
Quests	1/11, 2/11	Gaming	Recreational	625	1,072				

Medium	Source	Dates Messages Written (or forum posts started)	Dominant Topic	Intended Purpose	Desired Total Words to Collect	Collected		
						Total Words Per Forum	Total Words Per Source	Total Words Per Medium
Email	A multiple sclerosis support group	2/11 - 3/11	Other	Serious	3,750		3,747	
	Transcriptionists	2/11 - 3/11	Other	Serious	3,750		4,308	
	Fans of an opera singer	3/11	Other	Recreational	3,750		3,984	
	Beekeepers	2/11 - 3/11	Other	Ambiguous	3,750		4,636	
	Computing experts	3/11	Technology	Serious	7,500		7,508	
SMS	Multiplayer game players	10/10, 11/10, 12/10, 1/11 - 3/11	Gaming	Recreational	7,500		7,676	
	Dr. Susana Sotillo, Montclair University	2011	Other	Mixed	30,000		10,918	10,918
IM	UNC at Chapel Hill: Ask a Librarian (http://www.lib.unc.edu/ask.html)	5/07 - 5/08	Other	Serious	10,000		10,284	
	NCKnows, NC (http://ncknows.org/about.htm)	1/05 - 12/05	Other	Mixed	10,000		10,590	30,954
	L-Net, OR (http://www.oregonlibraries.net)	8/10 - 2/11	Other	Mixed	10,000		10,080	
Chat	World of Warcraft Chat (in-game)	3/09 - 4/09, 10/10 - 3/11	Gaming	Recreational	15,000		15,038	
	AOL Chat (http://chat.aim.com)	5/09 - 7/09	Other	Recreational	7,500		7,557	30,161
	NPS Chat Corpus (http://faculty.nps.edu/cmartell/NPSChat.htm)	2006	Other	Recreational	7,500		7,566	
Total Words in Corpus								136,529

Table 13 provides a higher-level, medium-centric view of the corpus. It shows the number of words desired for each medium, the number of individual messages collected, the number of words collected, and the average number of words per message.

Table 13: Word counts for media sections of the corpus

Medium	Proposed N Words to Collect	N Individual Messages Collected	N Words Collected	Average Words Per Message
Forums	30,000	835	32,637	39
Email Lists	30,000	412	31,859	77
SMS	30,000	1,391	10,918	8
IM	30,001	3,072	30,954	10
Chat	30,000	5,201	30,161	6
Totals	150,001	10,911	136,529	13

Table 14 shows the actual number of words collected for each of the topics and purposes.

Table 14: Word counts for topics and purposes

Topic					
Gaming and Technology	Gaming			31,379	47,096
	Technology	Technology	11,695	15,717	
		Gaming Technology	4,022		
Other				89,433	
Total				136,529	
Purpose					
Recreational				57,882	
Non-recreational	Serious			30,364	39,187
	Ambiguous			8,823	
Mixed				39,460	
Total				136,529	

Analysis

After cleaning the corpus, Perl scripts were used to create word frequency lists. “Corpus-linguistic analyses are always based on the evaluation of some kind of frequencies”: whether an individual element exists in the corpus, whether an element is more frequent than another element, or whether the observed frequency is more than what you would expect by chance (Gries, 2009, p. 1226). Kilgarriff (1997) outlines several advantages to using word frequency lists: they are (a) useful for text categorization, (b) susceptible to statistical processing, and (c) better for making similarity judgments than assessing the full text. “Any difference in the linguistic character of two corpora will leave its trace in a difference between their word frequency lists” (Kilgarriff, 1997, p. 233). This study examines the differences in feature frequency between different sections of the corpus—or subcorpora—for the purpose of uncovering any associations between those features and the medium or other situational variables.

The Perl scripts used in this study counted the number of times a unique word type appeared in each text and then output that data to a tab-delimited file that listed the term and its frequencies within each source (and each forum for the forum sources). For example, the term *lol* was found 151 times in AOL chat, 88 in NPS chat, 106 in WoW chat, once in the EverQuest Newbie Zone forum, once in the Teenspot General forum, and 39 in SMS. This list resulted in 23,912 unique word types.

This dissertation study combines automated (e.g., the Perl scripts used to create word frequency lists) and manual methods (e.g., classification of individual terms by features). Ball (1994, p. 295) explains “that given the present state of the art, automated methods and manual methods for text analysis must go hand in hand.” Automated methods should

“augment, but not replace, the human analysis process” (Ball, 1994, p. 296). Computational tools may be “imperfect” and so reliance upon them “may constrain the analysis process in undesirable ways” (Ball, 1994, p. 296). Ball (1994) explains that the danger in using purely automated methods lies in the possibility that the computer may miss things that only the human eye can detect. For example, criticism of Biber’s automated methods in his dissertation work suggests that Biber may have erroneously attributed effects to communicative function when they were really a matter of grammar (Grieve-Smith, 2007). Ball (1994) discusses how Biber did not account for every possible grammatical structure (e.g. zero complementizers),³⁸ and so it is possible he had confounding factors in his work.

In this dissertation study, standard/general English terms were stripped from the initial list of 23,912 word types—and saved to a separate file—so that only cyberlanguage candidate terms remained. This was done manually for the reasons outlined by Ball. An early experiment using an algorithm that removed terms based on their co-location in the word list and in a general English lexicon (e.g., WordNet) resulted in the removal of viable cyberlanguage candidates. For example, the term *pots* is a general English term often referring to cooking implements or containers for plants. However, in WoW chat, this term is a shortening for the word *potions* and as such is a viable cyberlanguage candidate that should not be removed from the word list.

Determinations about what terms should be removed were made based on intuition for obvious words (e.g., the, and, hamburger) and quick consultation with dictionaries.

³⁸ A *zero unit* in language is “postulated by an analysis, but which has no physical realization in the stream of speech” (Crystal, 2008a, p. 528). For example, “*Bob happy!*” does not include the verb *is* yet it is implied in the statement. *Is* becomes null or zero. Ball makes specific mention of zero complementizers and says that Biber wrote patterns that were meant to account for an unstated, yet implied (i.e., zero) *that*. Ball believes that Biber’s algorithms did not account for possible prepositional phrases or parentheticals that might serve in a zero capacity in the statements Biber was analyzing.

Questionable terms were retained because it was believed better to err on the side of including them at this stage. These questionable terms were reviewed later in more depth when each of the remaining terms was classified by the features it contained. If such terms later proved to be general English, they were removed at that time. However, it is possible that this quick initial pass risked eliminating viable cyberlanguage candidates. Ball (1994, p. 296) explains that manual analysis is not without risk either; it is “attended by tedium, errors, and the passage of time.” However, this quick initial pass was deemed to be a more practical approach to the removal of general English because manually reviewing the way each of the 24,000 word types was used in the corpus—i.e., each of the type’s tokens—would have been impractical in terms of time.

Slightly fewer than 14,000 terms were removed (13,990 specifically) in this initial weeding of standard/general English, leaving a substantially more manageable list of terms to process (9,924). Because at this time, there is no comprehensive, authoritative lexicon of cyberlanguage by which to make comparisons for quick and easy determination of term meanings and thus features used, manual classifications were required at this juncture as well. Thus, these remaining 9,924 terms were then manually analyzed in depth for the presence of the cyberlanguage features listed in Table 2. Each term was classified by one or more of these features. Any standard/general English terms that lingered in this list of 9,924 terms were also removed. To ensure classification results were not biased or inconsistent with feature definitions, an inter-coder reliability test was conducted that asked an outside coder to classify 5% of the terms by the features listed in Table 2.

Once all terms were classified, the frequencies with which features appeared in word tokens were compared via chi-square tests to determine if there were any significant

differences in feature use between media, media characteristics, sources, topics, or purposes. The rest of this section provides additional details about the classification process, the inter-coder reliability test, and the chi-square tests.

Classification

The classification process involved examining each cyberlanguage candidate term as it was used in the corpus, and determining what, if any, features in Table 2 might have been used in its creation. For example, the term *l2spell* (meaning *learn to spell*) would be classified as a single-letter form, number homophone, and a compound; *lol* (meaning *laughing out loud*) would be classified as an acronym.

The tab-delimited word list was opened in Microsoft Excel and classification decisions were made in this version of the file. Columns already existed for the unique word types and their frequencies. Columns were added for the terms' definitions and for the features listed in Table 2.

To classify a term, each instance of it was located in the corpus so that its usage could be examined. The surrounding context gave clues to term meaning, and once the meaning was determined, the features used became evident. Term meanings were entered into the Definition column, and *x* marks were placed in the column cells that matched the features the term used.

If term meaning was difficult to ascertain, then a term was researched on the Web. For example, if the meaning of a term used by the transcriptionist Google Group was not clear, other conversations from that group and any information about them on the Web were consulted. Sometimes UrbanDictionary.com was consulted or a general search of the term

using Google was conducted to help determine term meaning if not readily apparent from the context. Terms were also verified against two dictionaries available online:

- Merriam-Webster available on the Internet (a basic, general dictionary), and
- The Oxford English Dictionary available through the University Library's subscription (a more expansive dictionary).

If a term appeared in one or both of these dictionaries and was used in a manner consistent with its dictionary description, it was weeded from the list of cyberlanguage candidates and considered to be a left-over standard/general English term.

Initially, a little over 25% (2,590) of the 9,924 cyberlanguage candidate terms were classified by the researcher to clarify feature definitions and firm up classification rules in preparation for the inter-coder reliability test and to ensure consistent and reliable classification throughout analysis. This initial coding allowed the researcher to estimate the time the coder would need for classification and to refine coding rules and definitions shown in Table 2 as necessary. During classification of these 2,590 terms, three new features emerged and were added to the list of features:

- A State Abbreviations feature was added because many interlocutors referenced state abbreviations and may have done so in non-standard ways, such as not capitalizing these proper nouns.
- A Spelling Aloud feature was added because a word was spelled by separating each letter with a space, as if to signal pronunciation of each letter instead of pronouncing the entire word—e.g., *I want Y O U!*

- Another feature—Formatting Workarounds—was also added for instances where interlocutors invented ways to add special formatting that was not supported by the medium (e.g., bullets, footnotes, etc.).

The affixation feature was also expanded to include instances where combining forms were added to word bases. See Appendix D: Coding Rules and Appendix E: Signs and Symbols for the final list of features, their definitions, and their coding rules.

Inter-coder reliability test

Because manual classifications by one coder risk introducing bias into the analysis, an inter-coder reliability test on 5% of the cyberlanguage candidate terms was conducted. This test compared the classifications made by the researcher against those made by an outside coder. The goal was to ensure consistency and to confirm the reliability of the researcher's classifications.

Five percent was selected primarily for feasibility issues, i.e., manually coding terms is time-consuming and fatiguing. When classifying the initial 2,590 terms, the researcher took, on average, 1.5 hours to classify 100 terms. An outside coder who is less familiar with the corpus and the process of lexical analysis was estimated to take anywhere between 2.5 to 3 hours to classify 100 terms. Out of these 2,590 terms, 33% (845 terms) were lingering standard/general English and were removed. Based on this rate of attrition, the researcher estimated that once all 9,924 terms were examined, roughly 33% (3,275 terms) would be deemed standard/general English and be thus removed, leaving approximately 6,649³⁹ viable cyberlanguage terms. Five percent of 6,649 is 332, but this figure was rounded up to an even

³⁹ This estimate was very close to what actually remained—6,604 unique word types—after all word types were analyzed.

350 terms; 350 terms requires a less fatiguing and time-consuming effort from the coder than what 10 or more percent would require, and attention to detail and quality performance from the coder was desired.

The coder was selected for his familiarity with online communication media and games. The feature definitions and classification rules that were refined as a result of the researcher's initial classification of the approximately 25% of the 9,924 terms (2,590) were given to the coder, along with some general instructions on how to classify terms. (See Appendix D: Coding Rules and Appendix E: Signs and Symbols.) He was asked to first code a small training set of terms. These terms were not selected at random but were, instead, chosen because they exhibited at least one or more of the features listed in Table 2 as well as the new features uncovered during the researcher's initial coding of 2,590 terms. The goal of this training round was to give the coder an opportunity to become familiar with the features, to practice classifying terms, and to discuss his classification decisions with the researcher prior to classifying the production set of terms—i.e., the 5% (350 terms). Once the training round was complete and the coder felt comfortable with the process, he was asked to classify the 350 terms that comprised the production set. These terms were selected at random by the researcher, using a random number formula available in Microsoft Excel. The coder classified both the training set and the production set using an Excel spreadsheet similar to the one used by the researcher.

For each term the coder was asked to classify, he was allowed and encouraged to use three types of look-up sources to verify term meaning and thus determine feature use. The primary look-up source was the corpus itself, so the coder was given context examples—i.e.,

examples of each term's use in the corpus. The secondary look-up sources were Merriam-Webster and the Oxford English Dictionary.

The third look-up source was a general search of the term on the Web using Google. This source was to be used only in the event that the context examples and the online dictionaries proved unhelpful. Only a few terms required look-up on the general Web and usually this was because the term was specific to a particular communication situation and thus required some background knowledge of the situation that the coder did not possess without the Web research.

The classification of the production set was divided into three stages. At each stage the coder was asked to classify a third of the terms. Then when the coder finished, the researcher compared the coder's classifications to her own and calculated percent agreement. The researcher identified terms where both she and the coder disagreed, and then made the coder aware of these terms to give him an opportunity to change his classification decisions if, after a re-examination of them, he so desired. In identifying these terms for the coder, the researcher did not suggest or lead the coder to her own classification. If ultimately he believed his classification was correct, then it was not revised and the disagreement was factored into the later kappa calculations.

When the coder finished classifying all 350 terms, agreement was calculated using Artstein and Poesio's (2008) revised kappa, which was designed to avoid the prevalence paradox to which lexical analysis of this sort is subject. The prevalence paradox is a situation where most classifications will be of one kind, which would lead commonly-used inter-coder agreement calculations—such as Cohen's Kappa and Krippendorff's Alpha—to suggest that any agreements were purely chance. These algorithms are overly sensitive to these "rare"

categories. Artstein and Poesio (2008, p. 573) describe the prevalence paradox as a situation where when “data are highly skewed, coders may agree on a high proportion of items while producing annotations that are indeed correct to a high degree, yet the reliability coefficients remain low.” For example, in examining agreement for single-letter forms, only a few of the 350 terms will have exhibited this feature. This means that, if classifications are correct, the coder and the researcher will have marked only these few terms as exhibiting this feature—i.e., giving such terms a 1 rating, while the majority of terms would have been marked with 0s. Commonly-used agreement calculations will suggest low agreement because of this preponderance of 0s, even if the coder and researcher classified all 350 terms identically. The absence of more variation in 1s and 0s is not an indication of disagreement about this kind of data, because in actuality few terms indeed—out of a large number of terms—will exhibit any one feature. Realistically, it is highly improbable that any given term would exhibit even close to half of the features listed in Table 2.

Artstein and Poesio’s (2008) revised kappa is specifically designed for linguistic analyses in cases matching this dissertation study. It measures the “coders’ ability to agree on the rare category” (Artstein & Poesio, 2008, p. 573).

Artstein and Poesio (2008) recommend using Krippendorff’s (2004) schema for kappa interpretation, which Krippendorff drew from Carletta et al. (1997). Kappas greater than 0.80 demonstrate good reliability. Kappas between 0.67 and 0.80 allow for tentative conclusions, and kappas below 0.67 do not demonstrate reliability. Table 15 below shows Artstein and Poesio’s (2008) revised kappas for each feature category. Kappas in this table were rounded to the second decimal point, and cells with kappas that are below 0.80 are

highlighted in italics. Features with blank cells for the kappas did not appear in the production set.

Table 15: Inter-coder reliability kappas.

Feature Category	Kappa
Acronyms / initialisms	0.98
Shortenings	0.91
Clippings	1.00
Single-letter forms	0.90
Letter homophones	0.80
Number homophones	1.00
Symbolic substitution	1.00
Conjunctions	0.92
Disjunctions	<i>0.75</i>
Punctuation omission	1.00
Non-standard use of lowercase	0.96
State abbreviations	1.00
Onomatopoeic expression	0.95
Phonetic respellings (including elisions)	0.95
Offsetting punctuation	1.00
All caps	1.00
Letter duplication	0.97
Punctuation duplication	1.00
Spelling aloud	1.00
Emoticons	1.00
Emotes	0.80
Pointing	1.00
Pictograms	
Misspellings / typos	0.90
Repairs	1.00
Addressivity	
Reduplication	1.00
Affixation / combining forms	<i>0.76</i>
Compounds / space omission	0.98
Blends	
Conversion	0.89
Formatting workarounds	1.00

Chi-square tests

Once all terms were classified and all remaining standard/general English terms were removed (leaving 6,604 word types), frequencies for individual features were calculated and compared across media and genre factors using nonparametric chi-square (χ^2) goodness of fit tests. For example, the number of acronyms in synchronous media was compared to the number of acronyms used in asynchronous media to determine if there is a significant difference in acronym usage in these media.⁴⁰ A comparison of this sort attempts to determine whether synchronous media tend to foster more acronyms than asynchronous media, as is suggested in prior research. Gries (2009, p. 1228) explains that there is an “assumption underlying most corpus-based analyses ... that formal differences reflect, or correspond to, functional differences” and that “different frequencies of (co-)occurrences of formal elements ... are assumed to reflect functional regularities,” which are “intended to perform a particular communicative function.”

Chi-square tests the relationships between frequencies to determine if frequencies differ significantly from each other (Oakes, 1998). If values are statistically significant, then “you can conclude that there is an underlying relationship between the variables that is the basis for the frequency distribution you observed” (Wildemuth, 2009, p. 349).

Specifically, chi-square measures the difference between observed frequencies and theoretical expected frequencies (Wildemuth, 2009). Thus, chi-square specifies the difference as:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

⁴⁰ For these comparisons, token counts, not types, were used. For example, the acronym *lol* might have appeared 200 times in synchronous sources and 80 in asynchronous sources. The 200 and 80 helped constitute the token counts for acronyms.

where O represents observed frequencies, and E represents expected frequencies. The significance level (p value) is calculated by determining the degrees of freedom (in one-way tables, 1 less than the number of categorical variables) and comparing the statistic against the chi-square distribution.

Typically in one-way tables, the expected frequencies are calculated by dividing the total number of cases by the number of categorical variables (sources, texts, or other parameters by which features are being compared). For example, if one were comparing frequencies of verbs in three different sources and the observed frequencies totaled 3,000, the expected counts for each source would be 1,000. This approach does not consider differences in overall word frequencies for each source. It assumes that each of the three sources contain the same number of word tokens overall and that higher frequencies of a verb are due to some association between the source and feature when, in fact, such an association may not truly exist. Higher frequencies of verbs in a source that has more word tokens than another source with fewer tokens and fewer verbs may simply be a case where more verbs were possible because there are simply more words. For linguistic analyses, Biber et al. (1998) and Kilgarriff (1997) suggest normalizing based on token counts for each source, text, or other categorical variable.

In this dissertation study, expected counts were calculated by multiplying the total observed tokens for a feature by the proportion of tokens for a particular categorical variable out of all tokens in the corpus, as specified by:

$$E_i = n_{ip} \times O_t$$

where E_i is the expected count for a particular category (either a media or genre factor); n_{ip} is the proportion of i tokens—total tokens for that media or genre factor—out of N total tokens (n_i/N); and O_i is the total observed tokens for a particular feature. This can be expressed as follows:

$$E_i = O_t \left(\frac{n_i}{N} \right)$$

For example, there were 18,245 total tokens once all standard/general English terms were removed; 11,413 of these appeared in synchronous sources (0.6255 or 63% of 18,245), 6,832 in asynchronous sources (0.3744 or 37%). There were 2,776 acronyms in the corpus. Multiplying 2,776 by 0.6255 returns an expected count for synchronous acronyms of 1,736.50; and multiplying 2,776 by 0.3744 returns an expected count for asynchronous acronyms of 1,039.49. Where E_i is the expected count for synchronous acronyms and E_j is the expected count for asynchronous acronyms, their calculation is as follows:

$$\begin{aligned} E_i &= 2776(11,413/18,245) = 2776 \times 0.6255 = 1736.50 \\ E_j &= 2776(6,832/18,245) = 2776 \times 0.3744 = 1039.49 \end{aligned}$$

Frequencies for all features in Table 2 as well as the three features mentioned in the Classification section were compared in 13 different ways. Eight of these 13 sets of tests focused on media factors and included the following comparisons of feature frequency:

Table 16: Media factor comparisons.

Comparison	of Feature Frequency between
Medium	Forums Email lists SMS IM Chat
Synchronicity	Synchronous (chat, IM) Asynchronous (forums, email, SMS)
Participant Scale	1:1 (SMS, IM) 1:N (email lists) N:N (forums, chat)
Persistence (both visibility and re-use)	Extended (forums, email lists, SMS) Limited (IM, chat)
Anonymity	Greater anonymity possible (forums, IM, chat) Less anonymity possible(email lists, SMS)
Message Length	Limited (chat, IM, SMS) Unlimited (forums, email lists)
Compositional Ease	Easy (forums, email) Difficult (SMS) Partially difficult (chat, IM)
Viewing Ease	Easy (forums, email lists, IM, chat) Difficult (SMS)

The values chosen for these categorical variables—e.g., that chat and IM are synchronous and forums, email, and SMS are asynchronous—represent the most common designation for each media characteristic as shown in Table 1. Features were not compared based on differences in level of privacy afforded by a medium because, in effect, none of the conversations sampled were private if the researcher was able to obtain and use them for research purposes.

Five of the 13 sets of chi-square tests focused on genre factors—topic and purpose—and included the following comparisons of feature frequency:

Table 17: Genre factor comparisons.

Comparison	of Feature Frequency between
Topic 1	Gaming Technology Gaming Technology Other
Topic 2	Gaming Technology and Gaming Technology Other
Purpose 1	Serious Recreational/Leisure-oriented Mixed Ambiguous
Purpose 2	Serious Recreational/Leisure-oriented
Purpose 3	Non-recreational (serious, ambiguous) Recreational/Leisure-oriented

The values for these variables—e.g., that WoW chat was focused on Gaming topics and is used for Recreational/Leisure-oriented purposes—was discussed in the Corpus Creation section and in Appendix C: Support for Topic and Purpose Classifications.

Conclusion and Limitations

The research described here attempts to compare aspects of the communication situation with the linguistic features employed by interlocutors to uncover any associations for the purpose of verifying assertions made about cyberlanguage in prior research. The corpus used for these comparisons contains a mix of conversations from forums, email lists, SMS, IM, and chat, which permits comparison of linguistic feature frequency across a variety of media characteristics such as synchronicity, participant scale, message persistence, anonymity, message length restrictions, and compositional and viewing ease. Texts were taken from sources that vary in terms of topic and purpose, so that comparisons of linguistic feature frequency could be made based on these genre factors. Represented topics include

Gaming, Technology, and non-gaming/non-technology or Other topics. Represented purposes fall into two main categories: Serious and Recreational/leisure-oriented topics, with an additional two less-specific categories of Mixed purposes (i.e., discussions possibly for both serious and recreational purposes) and Ambiguous purposes (i.e., discussions where purpose is not fully determinant).

Most of the limitations of this corpus result from feasibility issues with text collection. For example, all IM sources are virtual reference conversations from libraries, which results in a somewhat homogeneous section of the corpus. However, without the generosity of the libraries that supplied these texts, no IM sources would have been included in this corpus. Only one SMS source could be obtained and it contains fewer word tokens than desired for that portion of the corpus. Although the messages come from a variety of interlocutors, fewer terms overall result in less breadth of language. The researcher did not have access to multiple types of online game chat, only WoW. So terminology for this section of the corpus will include specialized terms used by this particular gaming community and thus may not demonstrate as much breadth of general gaming lingo as originally desired. However, other sections of the corpus also include specialized terminology (e.g., the multiple sclerosis email list, the computing email list), so there is a mix of specialized vocabularies. Furthermore, it is probably impossible to avoid collection of such vocabularies.

As might be expected with manually analyzing thousands of terms, not all terms were easily disambiguated, so in some cases, it was difficult to classify the term. These terms were marked as “unknown.” There were several reasons terms might have been marked as unknown. First, both sides of an SMS conversation were not always represented in the corpus

and messages tended to be short. This meant that less context was available for disambiguating terms, so the researcher was unable to determine the meaning and, thus, the features used for some of the terms in this section of the corpus. Second, some portions of the corpus were not collected personally by the researcher, which meant that the ways in which these corpora were prepared for analysis was specific to the corpus creator's needs and not that of the researcher for this dissertation study. For example, userids were scrubbed from the NPS chat corpus before receipt of it. Some words in the NPS chat corpus appear to be shortened userids; but without the userids by which to make a comparison, it was difficult to determine these terms' exact meanings. Finally, because chat logging always started in the middle of one or more conversations, some terms in the chat section of the corpus refer to concepts discussed prior to logging, thus making it difficult to determine term meaning and features used without the initial reference. To combat these "unknowns," the researcher called upon the services of the coder from the inter-coder reliability test to help her disambiguate them. With his help, most of these "unknowns" were resolved. Ultimately, the coder and the researcher were unable to define only 116 terms (1.76% of the 6,604 cyberlanguage word types that were left after all general/standard English terms were removed at the completion of the classification stage). Despite the difficulties in design, collection, and cleaning, this corpus represents a unique resource for the study of cyberlanguage.

A final limitation worth mentioning has to do with the chi-square tests. Oakes (1998) and Wildemuth (2009) explain that an important limitation of chi-square testing is that as frequency increases, so do the chances of getting a high chi-square statistic, which leads to a significant p value. The greater the N , the more likely one will see significant p values

suggesting “relationships that are not really meaningful” (Wildemuth, 2009, p. 350). Thus, for this study only very small p values—less than .01 or, preferably, less than .001—were used as the basis for labeling comparisons as significant and potentially meaningful.

Findings

Introduction

The corpus contains 136,529 word tokens (23,912 word types). Once all general/standard English was removed, 14,681 word tokens (6,604 word types) remained. Table 18 shows the frequencies and percentages for tokens and types that were collected, that were general/standard English, and that are cyberlanguage candidate terms.

Table 18: Counts for words collected, general/standard English, and cyberlanguage terms.

	Collected	General/Standard English		Cyberlanguage	
		n	%	n	%
Tokens	136,529	121,848	89.25%	14,681	10.75%
Types	23,912	17,308	72.38%	6,604	27.62%

The 14,681 cyberlanguage candidate terms exhibited one or more of the features in Table 2 as well as the three features that emerged during classification (discussed in the Methods chapter): state abbreviations, spelling aloud, and formatting workarounds. Once the 14,681 tokens were classified by the feature(s) they contained, 18,245 feature codings resulted. Table 19 shows these feature frequencies and their percentages of the 18,245 features.

Table 19: Feature frequency and percent.

Feature	N	%
Acronyms / initialisms	2776	15.22%
Non-standard use of lowercase	2606	14.28%
Shortenings	1616	8.86%
Punctuation duplication	1295	7.10%
All caps	1229	6.74%
Compounds / space omission	1031	5.65%
Punctuation omission	1020	5.59%
Misspellings / typos	855	4.69%
Phonetic respellings	812	4.45%
Single-letter forms	730	4.00%
Emoticons	723	3.96%
Onomatopoeic expression	678	3.72%
Letter homophones	469	2.57%
Symbolic Substitution	417	2.29%
Offsetting punctuation	363	1.99%
Letter duplication	325	1.78%
Emotes	275	1.51%
Clippings	259	1.42%
Conjunctions / disjunctions	241	1.32%
Affixation / combining forms	143	0.78%
Conversion	109	0.60%
Number homophones	102	0.56%
State abbreviations	51	0.28%
Formatting workarounds	33	0.18%
Reduplication	30	0.16%
Repairs	25	0.14%
Pointing	19	0.10%
Spelling aloud	5	0.03%
Addressivity	5	0.03%
Blends	2	0.01%
Pictograms	1	0.01%
Total	18,245	100.00%

The results from the chi-square tests are shown below. Even though chi-square was calculated by normalizing based on subcorpora size, differences or similarities in feature frequencies were interpreted with caution, and broad claims were avoided for two main

reasons. First, there is disparity in subcorpora sizes as a result of both the corpus creation process and dividing the corpus for the purposes of making comparisons between media and genre factors, and this disparity could be reflected in differences in feature frequencies. Second, some features are scarce in the corpus (e.g., pictograms); and it is unknown whether their scarcity is due to the particular sample or whether they simply appear infrequently in online language as a whole.

Table 20 shows all features, their frequency, their chi-square values, and significance levels. Features with dashes (--) in the chi-square column had expected values less than 5, making them inappropriate for chi-square testing. Table 20 shows that features are not equally likely to occur in each of these media and these differences may be due to characteristics of the medium used for communication.

Table 20: Comparison of features among the five media.

Feature	Forums		Email		SMS		IM		Chat		χ^2	Sig
	n	%	n	%	n	%	n	%	n	%		
Acronyms / initialisms	437	15.74%	457	16.46%	185	6.66%	174	6.27%	1523	54.86%	(4, N=2776) = 236.43, $p = 0.000$	**
Shortenings	259	16.03%	122	7.55%	221	13.68%	162	10.02%	852	52.72%	(4, N=1616) = 14.19, $p = 0.007$	*
Clippings	28	10.81%	6	2.32%	62	23.94%	13	5.02%	150	57.92%	(4, N=259) = 56.29, $p = 0.000$	**
Single-letter forms	50	6.85%	38	5.21%	238	32.60%	50	6.85%	354	48.49%	(4, N=730) = 305.92, $p = 0.000$	**
Letter homophones	25	5.33%	3	0.64%	211	44.99%	30	6.40%	200	42.64%	(4, N=469) = 494.31, $p = 0.000$	**
Number homophones	12	11.76%	0	0.00%	47	46.08%	3	2.94%	40	39.22%	(4, N=102) = 114.50, $p = 0.000$	**
Symbolic Substitution	144	34.53%	46	11.03%	20	4.80%	52	12.47%	155	37.17%	(4, N=417) = 143.40, $p = 0.000$	**
Conjunctions / disjunctions	102	42.32%	37	15.35%	4	1.66%	29	12.03%	69	28.63%	(4, N=241) = 173.42, $p = 0.000$	**
Punctuation omission	144	14.12%	34	3.33%	151	14.80%	86	8.43%	605	59.31%	(4, N=1020) = 66.35, $p = 0.000$	**
Non-standard use of lowercase	368	14.12%	130	4.99%	288	11.05%	430	16.50%	1390	53.34%	(4, N=2606) = 161.89, $p = 0.000$	**
State abbreviations	2	3.92%	0	0.00%	4	7.84%	11	21.57%	34	66.67%	(4, N=51) = 18.21, $p = 0.001$	*
Onomatopoeic expression	69	10.18%	43	6.34%	57	8.41%	94	13.86%	415	61.21%	(4, N=678) = 46.73, $p = 0.000$	**
Phonetic respellings	75	9.24%	21	2.59%	154	18.97%	24	2.96%	538	66.26%	(4, N=812) = 168.47, $p = 0.000$	**
Offsetting punctuation	18	4.96%	161	44.35%	2	0.55%	5	1.38%	177	48.76%	(4, N=363) = 518.30, $p = 0.000$	**
All caps	158	12.86%	48	3.91%	39	3.17%	106	8.62%	878	71.44%	(4, N=1229) = 226.47, $p = 0.000$	**
Letter duplication	28	8.62%	8	2.46%	41	12.62%	38	11.69%	210	64.62%	(4, N=325) = 38.02, $p = 0.000$	**
Punctuation duplication	239	18.46%	172	13.28%	89	6.87%	229	17.68%	566	43.71%	(4, N=1295) = 134.92, $p = 0.000$	**
Spelling aloud	0	0.00%	4	80.00%	0	0.00%	0	0.00%	1	20.00%	--	
Emoticons	123	17.01%	173	23.93%	155	21.44%	64	8.85%	208	28.77%	(4, N=723) = 265.90, $p = 0.000$	**
Emotes	9	3.27%	30	10.91%	35	12.73%	4	1.45%	197	71.64%	(4, N=275) = 67.35, $p = 0.000$	**
Pointing	2	10.53%	0	0.00%	0	0.00%	0	0.00%	17	89.47%	--	
Pictograms	1	100.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	--	

Feature	Forums		Email		SMS		IM		Chat		χ^2	Sig
	n	%	n	%	n	%	n	%	n	%		
Misspellings / typos	159	18.60%	107	12.51%	80	9.36%	152	17.78%	357	41.75%		
Repairs	0	0.00%	1	4.00%	0	0.00%	3	12.00%	21	84.00%	--	
Addressivity	0	0.00%	5	100.00%	0	0.00%	0	0.00%	0	0.00%	--	
Reduplication	1	3.33%	1	3.33%	8	26.67%	1	3.33%	19	63.33%	--	
Affixation / combining forms	32	22.38%	16	11.19%	2	1.40%	22	15.38%	71	49.65%	(4, N=143) = 22.44, p = 0.000	**
Compounds / space omission	244	23.67%	147	14.26%	149	14.45%	130	12.61%	361	35.01%	(4, N=1031) = 134.12, p = 0.000	**
Blends	2	100.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	--	
Conversion	16	14.68%	2	1.83%	1	0.92%	1	0.92%	89	81.65%	(4, N=109) = 46.71, p = 0.000	**
Formatting workarounds	4	12.12%	26	78.79%	0	0.00%	3	9.09%	0	0.00%	--	
Total	2751	15.08%	1838	10.07%	2243	12.29%	1916	10.50%	9497	52.05%		

** p < .001

* p < .01

Table 20 provides a very general overview; so to determine which specific media and genre factors influence the production of certain features, comparisons of feature frequencies at more granular levels are shown in Tables 21 through 31.⁴¹ Each of these tables is discussed more fully in the sections that follow and the discussion will focus on those comparisons that resulted in significant chi-square values. A summary of major findings and qualitative details about specific features will be discussed in the Discussion chapter.

Statements comparing raw frequencies shown in the following tables may be misleading, so proportions of a feature's tokens out of all cyberlanguage tokens for a particular media or genre variable are used as the basis for determining where differences lie. This aligns with the method used for calculating expected values in the chi-square tests, and is thus more informative than using raw frequencies to sift out differences.

Synchronicity and Persistence

Table 21 below shows the comparison of feature frequency between synchronous, limited persistence media (chat and IM) and asynchronous, extended persistence media (forums, email, and SMS).

⁴¹ A Note about Predictive Texting as it Relates to the Interpretation of Tables 20 through 31: Predictive texting, available on some mobile devices, “automatically anticipates later letters in a word, based upon the initial letters entered” (Ling & Baron, 2013, p. 203). Although predictive texting may exert an influence on typing speed and terminology used (i.e., that someone could intend to enter a shortened form, but the text might be expanded), the researcher cannot make assertions about its possible effects, because she did not have data on the types or capabilities of mobile devices used by the participants in the SMS section of the corpus, or if predictive texting increased or hampered texting speed.

Table 21: Comparison of features between synchronous and asynchronous media.

Feature	Synchronous / Limited Persistence (chat, IM)		Asynchronous / Extended Persistence (forums, email, SMS)		χ^2	Sig
	n	%	n	%		
Acronyms / initialisms	1697	61.13%	1079	38.87%	(1, N=2776) = 2.40, $p = 0.121$	
Shortenings	1014	62.75%	602	37.25%	(1, N=1616) = 0.03, $p = 0.872$	
Clippings	163	62.93%	96	37.07%	(1, N=259) = 0.02, $p = 0.899$	
Single-letter forms	404	55.34%	326	44.66%	(1, N=730) = 16.21, $p = 0.000$	**
Letter homophones	230	49.04%	239	50.96%	(1, N=469) = 36.56, $p = 0.000$	**
Number homophones	43	42.16%	59	57.84%	(1, N=102) = 18.12, $p = 0.000$	**
Symbolic Substitution	207	49.64%	210	50.36%	(1, N=417) = 29.69, $p = 0.000$	**
Conjunctions / disjunctions	98	40.66%	143	59.34%	(1, N=241) = 49.30, $p = 0.000$	**
Punctuation omission	691	67.75%	329	32.25%	(1, N=1020) = 11.73, $p = 0.001$	*
Non-standard use of lowercase	1820	69.84%	786	30.16%	(1, N=2606) = 59.04, $p = 0.000$	**
State abbreviations	45	88.24%	6	11.76%	(1, N=51) = 14.36, $p = 0.000$	**
Onomatopoeic expression	509	75.07%	169	24.93%	(1, N=678) = 45.37, $p = 0.000$	**
Phonetic respellings	562	69.21%	250	30.79%	(1, N=812) = 15.37, $p = 0.000$	**
Offsetting punctuation	182	50.14%	181	49.86%	(1, N=363) = 23.89, $p = 0.000$	**
All caps	984	80.07%	245	19.93%	(1, N=1229) = 160.88, $p = 0.000$	**
Letter duplication	248	76.31%	77	23.69%	(1, N=325) = 26.25, $p = 0.000$	**
Punctuation duplication	795	61.39%	500	38.61%	(1, N=1295) = 0.75, $p = 0.387$	
Spelling aloud	1	20.00%	4	80.00%	--	
Emoticons	272	37.62%	451	62.38%	(1, N=723) = 16.21, $p = 0.000$	**
Emotes	201	73.09%	74	26.91%	(1, N=275) = 13.03, $p = 0.000$	**
Pointing	17	89.47%	2	10.53%	(1, N=19) = 5.88, $p = 0.015$	
Pictograms	0	0.00%	1	100.00%	--	
Misspellings / typos	509	59.53%	346	40.47%	(1, N=855) = 3.33, $p = 0.068$	
Repairs	24	96.00%	1	4.00%	(1, N=25) = 11.94, $p = 0.001$	*
Addressivity	0	0.00%	5	100.00%	--	
Reduplication	20	66.67%	10	33.33%	(1, N=30) = 0.22, $p = 0.642$	
Affixation / combining forms	93	65.03%	50	34.97%	(1, N=143) = 0.38, $p = 0.540$	
Compounds / space omission	491	47.62%	540	52.38%	(1, N=1031) = 98.12, $p = 0.000$	**
Blends	0	0.00%	2	100.00%	--	
Conversion	90	82.57%	19	17.43%	(1, N=109) = 18.64, $p = 0.000$	**
Formatting workarounds	3	9.09%	30	90.91%	(1, N=33) = 40.27, $p = 0.000$	**
Total	11413	62.55%	6832	37.45%		
** $p < .001$						
* $p < .01$						

Media classified as being synchronous were also classified as having limited persistence, and media classified as being asynchronous were also classified as having extended persistence. Consequently, there are some similarities in these two sets of media characteristics. For example, face-to-face speech is synchronous and, typically, non-persistent (i.e., unless someone is recording the speech). The text communication in synchronous, limited persistence media might then be thought of as more akin to speech than writing, as Hård af Segerstad (2002) and Davis and Brewer (1997) claim, for example. Thus, one might expect to find surrogate face-to-face cues in higher proportions in synchronous, limited persistence media than in asynchronous, extended persistence media, and the results in Table 21 support this. Onomatopoeic expression, phonetic respellings, all caps to convey shouting or emphasis, letter duplication, and emotes are some of the ways interlocutors may attempt to inject their own presence into the conversation and move it more toward a speech-like experience.

Furthermore, both synchronous and limited persistence media, like face-to-face speech, do not always allow interlocutors to look back on earlier utterances in the way interlocutors can when revising and editing a more persistent piece of writing. Thus, disfluencies, a natural part of speech often edited out in more traditional forms of writing, might also be expected to appear in higher proportions in synchronous, limited persistence media. However, differences in frequencies of misspellings and typos were not significant in these comparisons, which suggest that persistence and synchronicity have little influence over the production of errors; but the difference in frequency of repairs was significant. Repairs appeared in greater proportions in synchronous, limited persistence media. This

lends further support to Hård af Segerstad's (2002) and Davis and Brewer's (1997) assertions about the speech-like quality of synchronous, limited persistence media. Unplanned discourse in face-to-face conversations may require more mid-conversation or in-situ repairs than in writing, which is often planned. Johnstone (2008) explains that there is a greater preponderance of repair mechanisms in relatively unplanned discourse as opposed to relatively planned discourse.

Herring (2002) explains that synchronous communication may lead to more phatic communication and these findings—specifically the higher proportions of onomatopoeic expression, phonetic respellings, all caps, letter duplication, and emotes—may support Herring's assertion. Only one feature thought to convey emotion and sociality—emoticons—was found among those that are significant in asynchronous, extended persistence media. Additionally, emotes are performative, and allow interlocutors to communicate social and emotive information as Herring (2002) explains interlocutors are likely to do in synchronous settings. One of the motivations in using these face-to-face surrogates may be to convey a sense of one's self and bridge the distance inherent in online communication.

Punctuation omission, non-standard use of lowercase, onomatopoeic expression, phonetic respellings, all caps, letter duplication, emotes, repairs, and conversion appeared in higher proportions in synchronous, limited persistence media. Synchronous media are thought to encourage interlocutors to be brief (Ferrara et al., 1991; Herring, 2002), and Herring (2007) claims that low visibility persistence may also increase the likelihood of abbreviations. Yet there are more high-frequency abbreviations (5 to be exact) with significant chi-square values in asynchronous, extended persistence media than in synchronous, limited persistence media (2, or 3 if phonetic respellings also reduce

keystrokes). Furthermore, many of the high proportion features in synchronous, limited persistence media—such as letter duplication, all caps, onomatopoeic expression, and emotes—do not save time or keystrokes, and therefore would not result in shorter messages. This suggests that synchronous, limited persistence media do not induce greater abbreviation and brevity than asynchronous, extended persistence media.

Asynchronous, extended persistence media would seem to allow interlocutors more time to plan, edit, and review messages before sending. Thus, Crystal (2006, p. 140) claims that extended persistence media may push language more toward that seen in “articles, books, and other ‘permanent’ literature.” “There is an autonomy about the text, once it is posted, much like that encountered in a book” (Crystal, 2006, p. 140). Because messages from extended persistence media may reside longer on interlocutors’ computers and may be shared with others, one might expect to see more adherence to standard rules of grammar and spelling because such persistent communication may act as record of one’s conduct. Conversely, synchronous interactions are thought to cause the most “radical” linguistic innovations (Crystal, 2006, p. 135). When considering Crystal’s assertions, one would not expect to find many features in asynchronous, extended persistence media that bend the rules of typography and orthography. Yet in these data, several features that involve reshaping typography and orthography—single-letter forms, letter homophones, number homophones, offsetting punctuation, emoticons, compounds/space omission, and formatting workarounds—appear in higher proportions in asynchronous, extended persistence media, and their chi-square values are significant. Perhaps as interlocutors have more time to read and re-read messages, they are more likely to play and experiment with typography and orthography. Only two features were more frequent in asynchronous, extended persistence

media that are also found in general/standard English texts with some degree of frequency: conjunctions/disjunctions and symbolic substitution. (However, common uses of punctuation and other non-alphabetic symbols were excluded from consideration during analysis of terms in the corpus and were also higher in proportion in extended media. See Appendix E: Signs and Symbols.) So although there is a possibility that one's behavior can be tracked in extended persistence messages, this does not appear to be a motivation for avoiding typographical and orthographical deviations.

Furthermore, if one were to consider the use of a greater number of different kinds of cyberlanguage features to be evidence of greater innovation and less resemblance to traditional writing, Table 21—which shows roughly equal numbers of significant features in both synchronous, limited persistence and asynchronous, extended persistence media—would not support Crystal's assertion. Thus, probably the strongest link between synchronicity and persistence centers on the speech-like qualities of synchronous media and limited persistence media.

Participant Scale

Table 22 below shows the comparison of feature frequency between participant scales: 1:1 (one-to-one), 1:N (one-to-many), and N:N (many-to-many). Media were grouped into a participant scale category based on the most likely scale. For example, email was listed as 1:N because all email list discussions start off as one person writing an email to the entire group. Later, of course, the discussion may turn into a 1:1 (which is usually conducted off-list) or N:N conversation, but all discussions are built on the foundation of 1:N scales.

Table 22: Comparison of features by participant scale.

Feature	1:1 (SMS, IM)		1:N (email)		N:N (forums, chat)		χ^2	Sig
	n	%	n	%	n	%		
Acronyms / initialisms	359	12.93%	457	16.46%	1960	70.61%	(2, N=2776) = 235.92, p = 0.000	**
Shortenings	383	23.70%	122	7.55%	1111	68.75%	(2, N=1616) = 11.44, p = 0.003	*
Clippings	75	28.96%	6	2.32%	178	68.73%	(2, N=259) = 19.88, p = 0.000	**
Single-letter forms	288	39.45%	38	5.21%	404	55.34%	(2, N=730) = 121.14, p = 0.000	**
Letter homophones	241	51.39%	3	0.64%	225	47.97%	(2, N=469) = 235.26, p = 0.000	**
Number homophones	50	49.02%	0	0.00%	52	50.98%	(2, N=102) = 45.01, p = 0.000	**
Symbolic Substitution	72	17.27%	46	11.03%	299	71.70%	(2, N=417) = 7.27, p = 0.026	
Conjunctions / disjunctions	33	13.69%	37	15.35%	171	70.95%	(2, N=241) = 15.95, p = 0.000	**
Punctuation omission	237	23.24%	34	3.33%	749	73.43%	(2, N=1020) = 52.12, p = 0.000	**
Non-standard use of lowercase	718	27.55%	130	4.99%	1758	67.46%	(2, N=2606) = 92.81, p = 0.000	**
State abbreviations	15	29.41%	0	0.00%	36	70.59%	(2, N=51) = 6.21, p = 0.045	
Onomatopoeic expression	151	22.27%	43	6.34%	484	71.39%	(2, N=678) = 11.28, p = 0.004	*
Phonetic respellings	178	21.92%	21	2.59%	613	75.49%	(2, N=812) = 53.92, p = 0.000	**
Offsetting punctuation	7	1.93%	161	44.35%	195	53.72%	(2, N=363) = 502.47, p = 0.000	**
All caps	145	11.80%	48	3.91%	1036	84.30%	(2, N=1229) = 165.56, p = 0.000	**
Letter duplication	79	24.31%	8	2.46%	238	73.23%	(2, N=325) = 20.82, p = 0.000	**
Punctuation duplication	318	24.56%	172	13.28%	805	62.16%	(2, N=1295) = 19.75, p = 0.000	**
Spelling aloud	0	0.00%	4	80.00%	1	20.00%	--	
Emoticons	219	30.29%	173	23.93%	331	45.78%	(2, N=723) = 121.14, p = 0.000	**
Emotes	39	14.18%	30	10.91%	206	74.91%	(2, N=275) = 11.62, p = 0.003	*
Pointing	0	0.00%	0	0.00%	19	100.00%	--	
Pictograms	0	0.00%	0	0.00%	1	100.00%	--	
Misspellings / typos	232	27.13%	107	12.51%	516	60.35%	(2, N=855) = 17.97, p = 0.000	**
Repairs	3	12.00%	1	4.00%	21	84.00%	--	
Addressivity	0	0.00%	5	100.00%	0	0.00%	--	

Feature	1:1 (SMS, IM)		1:N (email)		N:N (forums, chat)		χ^2	Sig
	n	%	n	%	n	%		
Reduplication	9	30.00%	1	3.33%	20	66.67%	--	
Affixation / combining forms	24	16.78%	16	11.19%	103	72.03%	(2, N=143) = 2.95, p = 0.228	
Compounds / space omission	279	27.06%	147	14.26%	605	58.68%	(2, N=1031) = 37.11, p = 0.000	**
Blends	0	0.00%	0	0.00%	2	100.00%	--	
Conversion	2	1.83%	2	1.83%	105	96.33%	(2, N=109) = 42.20, p = 0.000	**
Formatting workarounds	3	9.09%	26	78.79%	4	12.12%	--	
Total	4159	22.80%	1838	10.07%	12248	67.13%		

** p < .001
 * p < .01

Werry (1996) explains that in many-to-many situations, interlocutors may feel pressure to keep up with the fast pace of the conversation by responding quickly, and “unless one can type very rapidly, messages must be kept short” (p. 53). Thus interlocutors may be more likely to abbreviate as the number of participants increases. Yet, in these data, the only abbreviation that appeared in higher proportions in N:N media and had a significant chi-square value was punctuation omission. (Phonetic respellings were also significant and appeared in higher proportions in N:N media. Sometimes these respellings result in shortened forms, and so those instances that do abbreviate could be considered along with punctuation omission. However, even when combining counts of phonetic respellings with punctuation omission, there are still fewer abbreviations in N:N media than in 1:1 or 1:N.) The other high-frequency features were onomatopoeic expression, all caps, letter duplication, emotes, and conversion. Most of these add characters and/or keystrokes.

Most significant abbreviations were higher in proportion in 1:1 media. Shortenings, clippings, single-letter forms, letter homophones, number homophones, and non-standard use of lowercase compose the list. Misspellings and typos were also higher in proportion for 1:1. This suggests that a fast pace (induced by N:N situations) does not necessarily cause interlocutors to make errors in their attempts to keep up, as one might assume.

Although N:N situations may not lead to higher proportions of abbreviations and errors, multiple-participant situations may lead interlocutors to create more surrogate face-to-face cues. The majority of the high proportion features with significant chi-square values shown for N:N situations are surrogates. Offsetting punctuation, punctuation duplication, and emoticons were in high proportions in 1:N situations, which may evolve into N:N conversations. Researchers such as Cherny (1999), Crystal (2006), Danet et al., (1997), and

Werry (1996) suggest that many-to-many media—particularly chat—may invite interlocutors to play with language and identity, and to create a highly interactive and performative experience with their fellow interlocutors. The high proportions of surrogate face-to-face cues with significant chi-square values support this idea. Onomatopoeic expression, emotes, and phonetic respellings—to name a few—are ways interlocutors can playfully inject sound and action into conversations, performing self through text.

Anonymity

Table 23 below shows the comparison of feature frequency between media with greater possibility for interlocutors to remain anonymous (forums, IM, chat) and media with lesser possibility for interlocutors to remain anonymous (email, SMS). Most features that are statistically significant are so at the less than 0.01 level, thus anonymity is not as strong a discriminator of feature use as some other media characteristics.

Table 23: Comparison of features by the degree of anonymity afforded by the medium.

Feature	Anonymity: Greater (forums, IM, chat)		Anonymity: Lesser (email, SMS)		χ^2	Sig
	n	%	n	%		
Acronyms / initialisms	2134	76.87%	642	23.13%	(1, N=2776) = 0.92, $p = 0.337$	
Shortenings	1273	78.77%	343	21.23%	(1, N=1616) = 1.21, $p = 0.270$	
Clippings	191	73.75%	68	26.25%	(1, N=259) = 2.25, $p = 0.133$	
Single-letter forms	454	62.19%	276	37.81%	(1, N=730) = 100.23, $p = 0.000$	**
Letter homophones	255	54.37%	214	45.63%	(1, N=469) = 146.14, $p = 0.000$	**
Number homophones	55	53.92%	47	46.08%	(1, N=102) = 33.02, $p = 0.000$	**
Symbolic Substitution	351	84.17%	66	15.83%	(1, N=417) = 10.27, $p = 0.001$	*
Conjunctions / disjunctions	200	82.99%	41	17.01%	(1, N=241) = 3.98, $p = 0.046$	
Punctuation omission	835	81.86%	185	18.14%	(1, N=1020) = 10.51, $p = 0.001$	*
Non-standard use of lowercase	2188	83.96%	418	16.04%	(1, N=2606) = 60.09, $p = 0.000$	**

Feature	Anonymity: Greater (forums, IM, chat)		Anonymity: Lesser (email, SMS)		χ^2	Sig
	n	%	n	%		
State abbreviations	47	92.16%	4	7.84%	(1, N=51) = 6.20, $p = 0.013$	
Onomatopoeic expression	578	85.25%	100	14.75%	(1, N=678) = 22.66, $p = 0.000$	**
Phonetic respellings	637	78.45%	175	21.55%	(1, N=812) = 0.31, $p = 0.577$	
Offsetting punctuation	200	55.10%	163	44.90%	(1, N=363) = 106.17, $p = 0.000$	**
All caps	1142	92.92%	87	7.08%	(1, N=1229) = 165.44, $p = 0.000$	**
Letter duplication	276	84.92%	49	15.08%	(1, N=325) = 9.95, $p = 0.002$	*
Punctuation duplication	1034	79.85%	261	20.15%	(1, N=1295) = 3.65, $p = 0.056$	
Spelling aloud	1	20.00%	4	80.00%	--	
Emoticons	395	54.63%	328	45.37%	(1, N=723) = 100.23, $p = 0.000$	**
Emotes	210	76.36%	65	23.64%	(1, N=275) = 0.25, $p = 0.614$	
Pointing	19	100.00%	0	0.00%	--	
Pictograms	1	100.00%	0	0.00%	--	
Misspellings / typos	668	78.13%	187	21.87%	(1, N=855) = 0.12, $p = 0.728$	
Repairs	24	96.00%	1	4.00%	(1, N=25) = 4.86, $p = 0.028$	
Addressivity	0	0.00%	5	100.00%	--	
Reduplication	21	70.00%	9	30.00%	(1, N=30) = 1.01, $p = 0.316$	
Affixation / combining forms	125	87.41%	18	12.59%	(1, N=143) = 7.88, $p = 0.005$	
Compounds / space omission	735	71.29%	296	28.71%	(1, N=1031) = 23.88, $p = 0.000$	**
Blends	2	100.00%	0	0.00%	--	
Conversion	106	97.25%	3	2.75%	(1, N=109) = 24.15, $p = 0.000$	**
Formatting workarounds	7	21.21%	26	78.79%	(1, N=33) = 60.49, $p = 0.000$	**
Total	14164	77.63%	4081	22.37%		
** $p < .001$						
* $p < .01$						

Crystal (2006) suggests that interlocutors may feel less inhibited in media that affords greater levels of anonymity. They can don a mask of their choosing and be whoever they wish to be online, making highly anonymous venues more like a masked ball (Danet, 2001). Crystal (2006, p. 54) claims that interlocutors “may feel emboldened to talk more and in different ways from their real-world linguistic repertoire.” Play with identity, language play, increased self-disclosure, and flaming may be more likely to occur in media that afford greater anonymity (Danet, 2001; Herring, 2007). Surrogate face-to-face cues are linguistic strategies

that interlocutors can employ to inject a sense of self into the conversation, and they could be viewed as playful and creative because they require the manipulation of typography and orthography to solve the problem of lack of face-to-face cues found in most cybermedia. In this corpus, however, only three surrogates appeared in higher proportions in media with greater possibility for anonymity: onomatopoeic expression, all caps, and letter duplication. In media with lesser possibility for anonymity, two high-proportion surrogates appeared: offsetting punctuation and emoticons. The difference in these numbers (three surrogates in media affording greater anonymity, two in media affording less anonymity) is minimal. So differences in anonymity may not cause interlocutors to use any more or less distinct surrogates.

The number of different types of abbreviation was also roughly equivalent—three were higher in proportion in media affording greater anonymity (symbolic substitution, punctuation omission, non-standard use of lowercase) and three were higher in proportion in media affording less anonymity (single-letter forms, letter homophones, number homophones). Thus anonymity may not lead to great diversity in the types of abbreviations or surrogates used, but a few specific ones are shown to differ. Overall this suggests that anonymity has little impact on one's use of cyberlanguage.

Message Length

Table 24 below shows the comparison of feature frequency between media with message length restrictions (chat, IM, SMS) and media without such limitations (email, forums).

Table 24: Comparison of features by message length restrictions.

Feature	Message Length: Limited (chat, IM, SMS)		Message Length: Unlimited (forums, email)		χ^2	Sig
	n	%	n	%		
Acronyms / initialisms	1882	67.80%	894	32.20%	(1, N=2776) = 73.34, $p = 0.000$	**
Shortenings	1235	76.42%	381	23.58%	(1, N=1616) = 2.13, $p = 0.144$	
Clippings	225	86.87%	34	13.13%	(1, N=259) = 19.89, $p = 0.000$	**
Single-letter forms	642	87.95%	88	12.05%	(1, N=730) = 66.52, $p = 0.000$	**
Letter homophones	441	94.03%	28	5.97%	(1, N=469) = 91.67, $p = 0.000$	**
Number homophones	90	88.24%	12	11.76%	(1, N=102) = 9.71, $p = 0.002$	*
Symbolic Substitution	227	54.44%	190	45.56%	(1, N=417) = 92.28, $p = 0.000$	**
Conjunctions / disjunctions	102	42.32%	139	57.68%	(1, N=241) = 135.42, $p = 0.000$	**
Punctuation omission	842	82.55%	178	17.45%	(1, N=1020) = 32.13, $p = 0.000$	**
Non-standard use of lowercase	2108	80.89%	498	19.11%	(1, N=2606) = 50.54, $p = 0.000$	**
State abbreviations	49	96.08%	2	3.92%	(1, N=51) = 12.21, $p = 0.000$	**
Onomatopoeic expression	566	83.48%	112	16.52%	(1, N=678) = 26.84, $p = 0.000$	**
Phonetic respellings	716	88.18%	96	11.82%	(1, N=812) = 76.63, $p = 0.000$	**
Offsetting punctuation	184	50.69%	179	49.31%	(1, N=363) = 112.54, $p = 0.000$	**
All caps	1023	83.24%	206	16.76%	(1, N=1229) = 45.96, $p = 0.000$	**
Letter duplication	289	88.92%	36	11.08%	(1, N=325) = 34.20, $p = 0.000$	**
Punctuation duplication	884	68.26%	411	31.74%	(1, N=1295) = 29.83, $p = 0.000$	**
Spelling aloud	1	20.00%	4	80.00%	--	
Emoticons	427	59.06%	296	40.94%	(1, N=723) = 66.52, $p = 0.000$	**
Emotes	236	85.82%	39	14.18%	(1, N=275) = 17.58, $p = 0.000$	**
Pointing	17	89.47%	2	10.53%	--	
Pictograms	0	0.00%	1	100.00%	--	
Misspellings / typos	589	68.89%	266	31.11%	(1, N=855) = 16.13, $p = 0.000$	**
Repairs	24	96.00%	1	4.00%	(1, N=25) = 5.94, $p = 0.015$	
Addressivity	0	0.00%	5	100.00%	--	
Reduplication	28	93.33%	2	6.67%	(1, N=30) = 5.45, $p = 0.020$	
Affixation / combining forms	95	66.43%	48	33.57%	(1, N=143) = 5.38, $p = 0.020$	
Compounds / space omission	640	62.08%	391	37.92%	(1, N=1031) = 89.34, $p = 0.000$	**
Blends	0	0.00%	2	100.00%	--	
Conversion	91	83.49%	18	16.51%	(1, N=109) = 4.32, $p = 0.038$	
Formatting workarounds	3	9.09%	30	90.91%	(1, N=33) = 75.80, $p = 0.000$	**
Total	13656	74.85%	4589	25.15%		
** $p < .001$						
* $p < .01$						

Ferrara et al. (1991), Herring (2007), Thurlow (2003), Werry (1996), and others suggest that the fewer characters allowed, the more likely interlocutors are to abbreviate. Six different types of abbreviation—clippings, single-letter forms, letter homophones, number homophones, punctuation omission, and non-standard use of lowercase—appeared in higher proportions in media with message length restrictions, while only three—acronyms, symbolic substitution, and conjunctions/disjunctions—appeared in higher proportions in media with few to no restrictions. This would appear to support the ideas of these other researchers; however, many character- or keystroke-adding features such as all caps, letter duplication, and emotes also appeared in higher proportions in media with limitations on message length. Hård af Segerstad (2002) and Thurlow (2003) claim that although interlocutors may be limited, their desire to establish group belonging and social identity through phatic communication may override the need to be brief. Phonetic respellings may help to achieve both goals of brevity and social belonging and they appear in higher proportions in media with limited message lengths. Some shorten words, yet they—like other surrogates—may also establish social presence, add humor, and lighten the tone of the conversation.

Features higher in proportion in media with extended message lengths include the three abbreviations mentioned previously, as well as offsetting punctuation, punctuation duplication, emoticons, misspellings/typos, compounds/space omission, and formatting workarounds. One might assume that there would be fewer misspellings and typos in media that impose little or no limitations on message length because more real estate would seem to give interlocutors more wiggle room for planning and editing messages, thus ensuring fewer errors. However, this is not the case in these data. Instead it would appear that when interlocutors are constrained by limited message lengths, they are more likely to be more

careful with their communication. With only so many characters in which to compose a message, every character may count heavily toward establishing clarity and understanding. Misspent characters could lead to confusion, and may later have to be repaired, which may exact greater costs to interlocutors wishing to be clear. Media with few to no restrictions on message length, however, allow for more elaboration; so perhaps interlocutors feel less concerned about making errors because the additional information that may be included in messages may help receivers resolve any ambiguities caused by errors.

Compositional Ease

Table 25 below shows the comparison of feature frequency between different levels of compositional ease. Compositional ease refers to any ergonomic difficulties one might encounter during message composition, such as small keyboard size (e.g., SMS) or small composition text boxes (e.g., chat and IM applications).

Table 25: Comparison of features by compositional ease.

Feature	Compositional Ease: Easy (forums, email)		Compositional Ease: Difficult (SMS)		Compositional Ease: Partially Difficult (chat, IM)		χ^2	Sig
	n	%	n	%	n	%		
Acronyms / initialisms	894	32.20%	185	6.66%	1697	61.13%	(2, N=2776) = 127.35, <i>p</i> = 0.000	**
Shortenings	381	23.58%	221	13.68%	1014	62.75%	(2, N=1616) = 4.11, <i>p</i> = 0.128	
Clippings	34	13.13%	62	23.94%	163	62.93%	(2, N=259) = 43.46, <i>p</i> = 0.000	**
Single-letter forms	88	12.05%	238	32.60%	404	55.34%	(2, N=730) = 300.77, <i>p</i> = 0.000	**
Letter homophones	28	5.97%	211	44.99%	230	49.04%	(2, N=469) = 490.12, <i>p</i> = 0.000	**
Number homophones	12	11.76%	47	46.08%	43	42.16%	(2, N=102) = 108.75, <i>p</i> = 0.000	**
Symbolic Substitution	190	45.56%	20	4.80%	207	49.64%	(2, N=417) = 99.26, <i>p</i> = 0.000	**
Conjunctions / disjunctions	139	57.68%	4	1.66%	98	40.66%	(2, N=241) = 141.99, <i>p</i> = 0.000	**
Punctuation omission	178	17.45%	151	14.80%	691	67.75%	(2, N=1020) = 33.67, <i>p</i> = 0.000	**
Non-standard use of lowercase	498	19.11%	288	11.05%	1820	69.84%	(2, N=2606) = 63.21, <i>p</i> = 0.000	**
State abbreviations	2	3.92%	4	7.84%	45	88.24%	(2, N=51) = 15.34, <i>p</i> = 0.000	**
Onomatopoeic expression	112	16.52%	57	8.41%	509	75.07%	(2, N=678) = 45.41, <i>p</i> = 0.000	**
Phonetic respellings	96	11.82%	154	18.97%	562	69.21%	(2, N=812) = 92.51, <i>p</i> = 0.000	**
Offsetting punctuation	179	49.31%	2	0.55%	182	50.14%	(2, N=363) = 133.90, <i>p</i> = 0.000	**
All caps	206	16.76%	39	3.17%	984	80.07%	(2, N=1229) = 177.80, <i>p</i> = 0.000	**
Letter duplication	36	11.08%	41	12.62%	248	76.31%	(2, N=325) = 35.45, <i>p</i> = 0.000	**
Punctuation duplication	411	31.74%	89	6.87%	795	61.39%	(2, N=1295) = 53.57, <i>p</i> = 0.000	**
Spelling aloud	4	80.00%	0	0.00%	1	20.00%	--	
Emoticons	296	40.94%	155	21.44%	272	37.62%	(2, N=723) = 300.77, <i>p</i> = 0.000	**
Emotes	39	14.18%	35	12.73%	201	73.09%	(2, N=275) = 18.08, <i>p</i> = 0.000	**
Pointing	2	10.53%	0	0.00%	17	89.47%	--	
Pictograms	1	100.00%	0	0.00%	0	0.00%	--	
Misspellings / typos	266	31.11%	80	9.36%	509	59.53%	(2, N=855) = 19.32, <i>p</i> = 0.000	**
Repairs	1	4.00%	0	0.00%	24	96.00%	--	

Feature	Compositional Ease: Easy (forums, email)		Compositional Ease: Difficult (SMS)		Compositional Ease: Partially Difficult (chat, IM)		χ^2	Sig
	n	%	n	%	n	%		
Addressivity	5	100.00%	0	0.00%	0	0.00%	--	
Reduplication	2	6.67%	8	26.67%	20	66.67%	--	
Affixation / combining forms	48	33.57%	2	1.40%	93	65.03%	(2, N=143) = 17.97, p = 0.000	**
Compounds / space omission	391	37.92%	149	14.45%	491	47.62%	(2, N=1031) = 107.52, p = 0.000	**
Blends	2	100.00%	0	0.00%	0	0.00%	--	
Conversion	18	16.51%	1	0.92%	90	82.57%	(2, N=109) = 21.69, p = 0.000	**
Formatting workarounds	30	90.91%	0	0.00%	3	9.09%	--	
Total	4589	25.15%	2243	12.29%	11413	62.55%		

** p < .001

* p < .01

The more difficult it is to compose a message, the more likely one might want to do it succinctly so that less time is spent on a task that is potentially frustrating. Thurlow's (2003) assessment of brevity in SMS—that it results from a need for speed and ease of typing—supports this idea. So one might expect to see more abbreviations in SMS, chat, and IM (those media with difficult or partially difficult compositional ease) than in forums and email (media that afford easy composition). Five different types of abbreviations (clippings, single-letter forms, letter homophones, number homophones, and punctuation omission) appeared in higher proportions in media with difficult composition, as opposed to the two (non-standard lowercase and state abbreviations) in partially difficult media and the three (acronyms, symbolic substitution, and conjunctions/disjunctions) in media with easy composition. Taken alone or with media with partially difficult composition, media with difficult composition exhibit more types of abbreviations, which supports the supposition made earlier about the need for speed and ease of typing. Phonetic respellings (which may also result in a shortened word) and emoticons were also more frequent in these media.

Typing punctuation on a cellphone's small keyboard or touch-screen keypad may be difficult. Apostrophes may require as many as four taps on a cellphone's keyboard as opposed to one keystroke on a full-sized computer keyboard (Ling & Baron, 2007). Baron (2008) fears the demise of apostrophes online, claiming that they may become an "endangered species" (p. 61). Thurlow (2003) believes they are not yet dead and in his 544-message SMS corpus, he found 192 examples of apostrophes. However, Ling and Baron's (2007) comparison of SMS and IM showed far fewer contractions using an apostrophe in SMS (31.9% of contractions) than in IM (93.9% of contractions). In this dissertation corpus, most types of punctuation omission were contractions without apostrophes, and they appear

in greater proportions in media with the greatest challenges to composition (i.e., SMS). This supports Baron's (2008) and Ling and Baron's (2007) findings. However, these data also show higher proportions of emoticons in media with the greatest challenges to composition, and all emoticons (except graphic emoticons) typically include one or more punctuation marks. If apostrophes were tedious to type, then it would follow that other punctuation marks would also be tedious to type. Yet these results show that interlocutors were willing to work through the difficulties where emoticons are concerned. Many emoticons do not include letters and instead consist of only punctuation (e.g., :-()) or punctuation and numerals (e.g., &-); and on many touch-screen mobile devices—for example, iPhones—one can type the full emoticon without stroking keys multiple times (as is required on non-touch-screen phones) or without switching back and forth between touch-screen keypads. Figure 2 shows the punctuation and numerals keypad on an iPhone, where one can compose a number of emoticons in a single keypad.



Figure 5: Punctuation and numerals keypad on an iPhone.

Perhaps more SMS utterances in this corpus were created by interlocutors with touch-screen cellphones as opposed to cellphones with mini keyboards, or perhaps not including an apostrophe in a contraction is becoming a conventional way to abbreviate. Furthermore, apostrophes by themselves do not contain as much information as emoticons do. They are purely grammatical devices, whereas emoticons may convey emotion, sentiment, or intention. Skimping on an apostrophe might not lessen the impact or clarity of a message, but skimping on an emoticon might.

Higher-proportion, significant features in media with partially difficult message composition include non-standard lowercase, onomatopoeic expression, all caps, letter duplication, emotes, and conversion. Those higher in proportion in media with easy message composition include acronyms, symbolic substitution, conjunctions/disjunctions, offsetting punctuation, punctuation duplication, misspellings/typos, affixation/combining forms, and compounds/space omission.

Viewing Ease

Table 26 below shows the comparison of feature frequency between different levels of viewing ease. Viewing ease refers to screen or window sizes when viewing messages, such as large screens or application windows frequently available in forums, email, IM, and chat as opposed to the tiny screens found on mobile devices.

Table 26: Comparison of features by viewing ease.

Feature	Viewing Ease: Easy (all but SMS)		Viewing Ease: Difficult (SMS)		χ^2	Sig
	n	%	n	%		
Acronyms / initialisms	2591	93.34%	185	6.66%	(1, N=2776) = 81.59, $p = 0.000$	**
Shortenings	1395	86.32%	221	13.68%	(1, N=1616) = 2.86, $p = 0.091$	
Clippings	197	76.06%	62	23.94%	(1, N=259) = 32.57, $p = 0.000$	**
Single-letter forms	492	67.40%	238	32.60%	(1, N=730) = 279.24, $p = 0.000$	**
Letter homophones	258	55.01%	211	44.99%	(1, N=469) = 464.98, $p = 0.000$	**
Number homophones	55	53.92%	47	46.08%	(1, N=102) = 107.98, $p = 0.000$	**
Symbolic Substitution	397	95.20%	20	4.80%	(1, N=417) = 21.74, $p = 0.000$	**
Conjunctions / disjunctions	237	98.34%	4	1.66%	(1, N=241) = 25.28, $p = 0.000$	**
Punctuation omission	869	85.20%	151	14.80%	(1, N=1020) = 5.96, $p = 0.015$	
Non-standard use of lowercase	2318	88.95%	288	11.05%	(1, N=2606) = 3.73, $p = 0.053$	
State abbreviations	47	92.16%	4	7.84%	(1, N=51) = 0.94, $p = 0.333$	
Onomatopoeic expression	621	91.59%	57	8.41%	(1, N=678) = 9.50, $p = 0.002$	*
Phonetic respellings	658	81.03%	154	18.97%	(1, N=812) = 33.52, $p = 0.000$	**
Offsetting punctuation	361	99.45%	2	0.55%	(1, N=363) = 46.42, $p = 0.000$	**
All caps	1190	96.83%	39	3.17%	(1, N=1229) = 94.81, $p = 0.000$	**
Letter duplication	284	87.38%	41	12.62%	(1, N=325) = 0.03, $p = 0.860$	
Punctuation duplication	1206	93.13%	89	6.87%	(1, N=1295) = 35.30, $p = 0.000$	**
Spelling aloud	5	100.00%	0	0.00%	--	
Emoticons	568	78.56%	155	21.44%	(1, N=723) = 279.24, $p = 0.000$	**
Emotes	240	87.27%	35	12.73%	(1, N=275) = 0.05, $p = 0.827$	
Pointing	19	100.00%	0	0.00%	--	
Pictograms	1	100.00%	0	0.00%	--	
Misspellings / typos	775	90.64%	80	9.36%	(1, N=855) = 6.84, $p = 0.009$	
Repairs	25	100.00%	0	0.00%	--	
Addressivity	5	100.00%	0	0.00%	--	
Reduplication	22	73.33%	8	26.67%	--	
Affixation / combining forms	141	98.60%	2	1.40%	(1, N=143) = 15.74, $p = 0.000$	**
Compounds / space omission	882	85.55%	149	14.45%	(1, N=1031) = 4.45, $p = 0.035$	
Blends	2	100.00%	0	0.00%	--	
Conversion	108	99.08%	1	0.92%	(1, N=109) = 13.08, $p = 0.000$	**
Formatting workarounds	33	100.00%	0	0.00%	--	
Total	16002	87.71%	2243	12.29%		
** $p < .001$						
* $p < .01$						

Acronyms/initialisms, symbolic substitution, conjunctions/disjunctions, onomatopoeic expression, all caps, affixation/combining forms, and conversion have significant chi-square values and are proportionally higher in media with easy viewing. Clippings, single-letter forms, letter homophones, number homophones, phonetic respellings, and emoticons have significant chi-square values and are higher in proportion in media with difficult viewing. More types of abbreviations appear in higher proportions in media with difficult viewing, so they appear to have some association with small screens. Perhaps in the way that people try to make themselves smaller in packed elevators, interlocutors try to make their utterances smaller when composing in tiny virtual spaces. The feeling of being closed in may lead interlocutors to compact their comments.

Topic

Table 27 below shows the comparison of feature frequency between different topics: Gaming, Technology, Gaming Technology, and Other.⁴²

⁴² Refer to Table 12 in the Methods section for how sources/media were classified by topic.

Table 27: Comparison of features by gaming, technology, gaming technology, and other topics.

Feature	Gaming		Technology		Gaming Technology		Other		χ^2	Sig
	n	%	n	%	n	%	n	%		
Acronyms / initialisms	1148	41.35%	295	10.63%	62	2.23%	1271	45.79%	(3, N=2776) = 607.02, p = 0.000	**
Shortenings	646	39.98%	43	2.66%	62	3.84%	865	53.53%	(3, N=1616) = 150.37, p = 0.000	**
Clippings	43	16.60%	4	1.54%	2	0.77%	210	81.08%	--	
Single-letter forms	183	25.07%	11	1.51%	2	0.27%	534	73.15%	(3, N=730) = 31.54, p = 0.000	**
Letter homophones	54	11.51%	0	0.00%	0	0.00%	415	88.49%	(3, N=469) = 115.42, p = 0.000	**
Number homophones	21	20.59%	0	0.00%	0	0.00%	81	79.41%	--	
Symbolic Substitution	151	36.21%	31	7.43%	16	3.84%	219	52.52%	(3, N=417) = 40.25, p = 0.000	**
Conjunctions / disjunctions	76	31.54%	25	10.37%	13	5.39%	127	52.70%	--	
Punctuation omission	348	34.12%	17	1.67%	20	1.96%	635	62.25%	(3, N=1020) = 25.32, p = 0.000	**
Non-standard use of lowercase	806	30.93%	61	2.34%	38	1.46%	1701	65.27%	(3, N=2606) = 23.43, p = 0.000	**
State abbreviations	4	7.84%	0	0.00%	0	0.00%	47	92.16%	--	
Onomatopoeic expression	197	29.06%	15	2.21%	6	0.88%	460	67.85%	(3, N=678) = 9.20, p = 0.027	
Phonetic respellings	177	21.80%	5	0.62%	9	1.11%	621	76.48%	(3, N=812) = 55.73, p = 0.000	**
Offsetting punctuation	57	15.70%	7	1.93%	2	0.55%	297	81.82%	(3, N=363) = 44.38, p = 0.000	**
All caps	173	14.08%	28	2.28%	11	0.90%	1017	82.75%	(3, N=1229) = 166.61, p = 0.000	**
Letter duplication	37	11.38%	3	0.92%	3	0.92%	282	86.77%	(3, N=325) = 66.94, p = 0.000	**
Punctuation duplication	272	21.00%	59	4.56%	13	1.00%	951	73.44%	(3, N=1295) = 46.58, p = 0.000	**
Spelling aloud	1	20.00%	0	0.00%	0	0.00%	4	80.00%	--	
Emoticons	160	22.13%	21	2.90%	7	0.97%	535	74.00%	(3, N=723) = 31.54, p = 0.000	**
Emotes	41	14.91%	0	0.00%	1	0.36%	233	84.73%	--	
Pointing	5	26.32%	0	0.00%	0	0.00%	14	73.68%	--	
Pictograms	1	100.00%	0	0.00%	0	0.00%	0	0.00%	--	
Misspellings / typos	215	25.15%	35	4.09%	14	1.64%	591	69.12%	(3, N=855) = 6.38, p = 0.095	
Repairs	10	40.00%	0	0.00%	0	0.00%	15	60.00%	--	
Addressivity	0	0.00%	3	60.00%	0	0.00%	2	40.00%	--	

Feature	Gaming		Technology		Gaming Technology		Other		χ^2	Sig
	n	%	n	%	n	%	n	%		
Reduplication	8	26.67%	0	0.00%	0	0.00%	22	73.33%	--	
Affixation / combining forms	51	35.66%	7	4.90%	7	4.90%	78	54.55%	--	
Compounds / space omission	304	29.49%	57	5.53%	19	1.84%	651	63.14%	(3, N=1031) = 6.36, <i>p</i> = 0.095	
Blends	1	50.00%	0	0.00%	0	0.00%	1	50.00%	--	
Conversion	86	78.90%	2	1.83%	9	8.26%	12	11.01%	--	
Formatting workarounds	12	36.36%	11	33.33%	0	0.00%	10	30.30%	--	
Total	5288	28.98%	740	4.06%	316	1.73%	11901	65.23%		

** *p* < .001
 * *p* < .01

Of the features with significant chi-square values in Table 27, punctuation omission and non-standard use of lowercase were higher in proportion in conversations focused on Gaming topics; acronyms were higher in proportion in conversations focused on Technology topics; shortenings and symbolic substitution were higher in proportion in Gaming Technology discussions; and single-letter forms, letter homophones, phonetic respellings, all caps, letter duplication, punctuation duplication, and emoticons were more frequent in conversations focused on Other topics.

Table 28 shows the same comparison as Table 27 except that counts for Technology and Gaming Technology have been merged into a larger Technology category.

Table 28: Comparison of features by gaming, technology/gaming technology, and other topics.

Feature	Gaming		Technology / Gaming Technology		Other		χ^2	Sig
	n	%	n	%	n	%		
Acronyms / initialisms	1148	41.35%	357	12.86%	1271	45.79%	(2, N=2776) = 547.37, <i>p</i> = 0.000	**
Shortenings	646	39.98%	105	6.50%	865	53.53%	(2, N=1616) = 102.70, <i>p</i> = 0.000	**
Clippings	43	16.60%	6	2.32%	210	81.08%	(2, N=259) = 29.07, <i>p</i> = 0.000	**
Single-letter forms	183	25.07%	13	1.78%	534	73.15%	(2, N=730) = 31.13, <i>p</i> = 0.000	**
Letter homophones	54	11.51%	0	0.00%	415	88.49%	(2, N=469) = 115.42, <i>p</i> = 0.000	**
Number homophones	21	20.59%	0	0.00%	81	79.41%	(2, N=102) = 11.53, <i>p</i> = 0.003	*
Symbolic Substitution	151	36.21%	47	11.27%	219	52.52%	(2, N=417) = 39.51, <i>p</i> = 0.000	**
Conjunctions / disjunctions	76	31.54%	38	15.77%	127	52.70%	(2, N=241) = 47.81, <i>p</i> = 0.000	**
Punctuation omission	348	34.12%	37	3.63%	635	62.25%	(2, N=1020) = 18.89, <i>p</i> = 0.000	**
Non-standard use of lowercase	806	30.93%	99	3.80%	1701	65.27%	(2, N=2606) = 21.22, <i>p</i> = 0.000	**
State abbreviations	4	7.84%	0	0.00%	47	92.16%	--	
Onomatopoeic expression	197	29.06%	21	3.10%	460	67.85%	(2, N=678) = 9.19, <i>p</i> = 0.010	
Phonetic respellings	177	21.80%	14	1.72%	621	76.48%	(2, N=812) = 55.38, <i>p</i> = 0.000	**
Offsetting punctuation	57	15.70%	9	2.48%	297	81.82%	(2, N=363) = 44.27, <i>p</i> = 0.000	**
All caps	173	14.08%	39	3.17%	1017	82.75%	(2, N=1229) = 166.58, <i>p</i> = 0.000	**
Letter duplication	37	11.38%	6	1.85%	282	86.77%	(2, N=325) = 66.57, <i>p</i> = 0.000	**
Punctuation duplication	272	21.00%	72	5.56%	951	73.44%	(2, N=1295) = 41.94, <i>p</i> = 0.000	**
Spelling aloud	1	20.00%	0	0.00%	4	80.00%	--	
Emoticons	160	22.13%	28	3.87%	535	74.00%	(2, N=723) = 31.13, <i>p</i> = 0.000	**
Emotes	41	14.91%	1	0.36%	233	84.73%	(2, N=275) = 48.80, <i>p</i> = 0.000	**
Pointing	5	26.32%	0	0.00%	14	73.68%	--	
Pictograms	1	100.00%	0	0.00%	0	0.00%	--	
Misspellings / typos	215	25.15%	49	5.73%	591	69.12%	(2, N=855) = 6.34, <i>p</i> = 0.042	
Repairs	10	40.00%	0	0.00%	15	60.00%	--	
Addressivity	0	0.00%	3	60.00%	2	40.00%	--	

Feature	Gaming		Technology / Gaming Technology		Other		χ^2	Sig
	n	%	n	%	n	%		
Reduplication	8	26.67%	0	0.00%	22	73.33%	--	
Affixation / combining forms	51	35.66%	14	9.79%	78	54.55%	(2, N=143) = 8.66, p = 0.013	
Compounds / space omission	304	29.49%	76	7.37%	651	63.14%	(2, N=1031) = 5.24, p = 0.073	
Blends	1	50.00%	0	0.00%	1	50.00%	--	
Conversion	86	78.90%	11	10.09%	12	11.01%	(2, N=109) = 146.32, p = 0.000	**
Formatting workarounds	12	36.36%	11	33.33%	10	30.30%	--	
Total	5288	28.98%	1056	5.79%	11901	65.23%		

*** p < .001
* p < .01

When Technology and Gaming Technology are merged, a few more features become significant and some proportions shift. Punctuation omission and non-standard lowercase are still higher in proportion in Gaming discussions, but shortenings (formerly higher in proportion in Gaming Technology discussions) and conversion are added to the list. This suggests that gamers are likely to shorten no matter if they discuss ordinary game topics or topics related to gaming technology.

In addition to acronyms, symbolic substitution (formerly higher in proportion in Gaming Technology discussions) and conjunctions/disjunctions are added to the list of high frequency features with significant chi-square values in Technology discussions. The popular belief that information technology professionals overload their vocabulary with acronyms may be true, or perhaps “insiders” or members of a “community of practice” assume each other’s knowledge of acronyms and are thus more comfortable using them.

In addition to the features in Table 27 that were significant and proportionally higher in conversations focused on Other topics, clippings, number homophones, and emotes also appear on the list. Based on this list, which includes several ways to manipulate the orthography and typography of language, language play appears to be more prevalent in conversations focused on Other topics. Gaming topics do not appear to incite as much play as one might assume, and perhaps this is because gamers have been shown to take their gameplay very seriously, treating it almost as if it were a job (Yee, 2006).

Purpose

Table 29 below shows the comparison of feature frequency between different purposes: Serious, Recreational/Leisure-oriented, Mixed, and Ambiguous.⁴³

⁴³ Refer to Table 12 in the Methods section for how sources/media were classified by purpose.

Table 29: Comparison of features by serious, recreational/leisure-oriented, mixed, and ambiguous purposes.

Feature	Serious		Recreational		Mixed		Ambiguous		χ^2	Sig
	n	%	n	%	n	%	n	%		
Acronyms / initialisms	413	14.88%	1905	68.62%	316	11.38%	142	5.12%	(3, N=2776) = 233.85, <i>p</i> = 0.000	**
Shortenings	142	8.79%	1080	66.83%	356	22.03%	38	2.35%	(3, N=1616) = 12.41, <i>p</i> = 0.006	*
Clippings	8	3.09%	163	62.93%	87	33.59%	1	0.39%	(3, N=259) = 37.30, <i>p</i> = 0.000	**
Single-letter forms	33	4.52%	401	54.93%	286	39.18%	10	1.37%	(3, N=730) = 144.99, <i>p</i> = 0.000	**
Letter homophones	18	3.84%	204	43.50%	246	52.45%	1	0.21%	(3, N=469) = 266.36, <i>p</i> = 0.000	**
Number homophones	6	5.88%	45	44.12%	51	50.00%	0	0.00%	--	
Symbolic Substitution	71	17.03%	256	61.39%	75	17.99%	15	3.60%	(3, N=417) = 16.14, <i>p</i> = 0.001	*
Conjunctions / disjunctions	57	23.65%	144	59.75%	31	12.86%	9	3.73%	(3, N=241) = 42.48, <i>p</i> = 0.000	**
Punctuation omission	87	8.53%	672	65.88%	245	24.02%	16	1.57%	(3, N=1020) = 15.23, <i>p</i> = 0.002	*
Non-standard use of lowercase	341	13.09%	1618	62.09%	593	22.76%	54	2.07%	(3, N=2606) = 13.93, <i>p</i> = 0.003	*
State abbreviations	4	7.84%	36	70.59%	11	21.57%	0	0.00%	--	
Onomatopoeic expression	74	10.91%	465	68.58%	123	18.14%	16	2.36%	(3, N=678) = 6.98, <i>p</i> = 0.072	
Phonetic respellings	26	3.20%	575	70.81%	199	24.51%	12	1.48%	(3, N=812) = 60.90, <i>p</i> = 0.000	**
Offsetting punctuation	12	3.31%	339	93.39%	8	2.20%	4	1.10%	(3, N=363) = 137.34, <i>p</i> = 0.000	**
All caps	69	5.61%	956	77.79%	188	15.30%	16	1.30%	(3, N=1229) = 105.70, <i>p</i> = 0.000	**
Letter duplication	23	7.08%	226	69.54%	73	22.46%	3	0.92%	(3, N=325) = 10.83, <i>p</i> = 0.013	
Punctuation duplication	202	15.60%	758	58.53%	288	22.24%	47	3.63%	(3, N=1295) = 30.19, <i>p</i> = 0.000	**
Spelling aloud	3	60.00%	2	40.00%	0	0.00%	0	0.00%	--	
Emoticons	132	18.26%	299	41.36%	262	36.24%	30	4.15%	(3, N=723) = 227.52, <i>p</i> = 0.000	**
Emotes	19	6.91%	214	77.82%	42	15.27%	0	0.00%	(3, N=275) = 26.01, <i>p</i> = 0.000	**
Pointing	0	0.00%	18	94.74%	1	5.26%	0	0.00%	--	
Pictograms	0	0.00%	1	100.00%	0	0.00%	0	0.00%	--	
Misspellings / typos	166	19.42%	461	53.92%	205	23.98%	23	2.69%	(3, N=855) = 64.34, <i>p</i> = 0.000	**
Repairs	1	4.00%	22	88.00%	2	8.00%	0	0.00%	--	
Addressivity	5	100.00%	0	0.00%	0	0.00%	0	0.00%	--	
Reduplication	2	6.67%	20	66.67%	8	26.67%	0	0.00%	--	
Affixation / combining forms	26	18.18%	101	70.63%	13	9.09%	3	2.10%	--	

Feature	Serious		Recreational		Mixed		Ambiguous		χ^2	Sig
	n	%	n	%	n	%	n	%		
Compounds / space omission	121	11.74%	584	56.64%	270	26.19%	56	5.43%	(3, N=1031) = 44.23, p = 0.000	**
Blends	0	0.00%	1	50.00%	1	50.00%	0	0.00%	--	
Conversion	2	1.83%	104	95.41%	3	2.75%	0	0.00%	--	
Formatting workarounds	9	27.27%	12	36.36%	3	9.09%	9	27.27%	--	
Total	2072	11.36%	11682	64.03%	3986	21.85%	505	2.77%		

** p < .001
* p < .01

With the exception of onomatopoeic expression, the features listed in Table 29 for which chi-square tests could be conducted are not equally likely to occur in each of these four purposes. Tables 30 and 31 below provide more discriminating views of which features are more likely to differ in frequency based on purpose.

Table 30 removes Mixed and Ambiguous categories from comparison, and compares frequencies from only sources that have serious or recreational/leisure-oriented purposes.

Table 30: Comparison of features by serious and recreational/leisure-oriented purposes.

Feature	Serious		Recreational		χ^2	Sig
	n	%	n	%		
Acronyms / initialisms	413	17.82%	1905	82.18%	(1, N=2318) = 13.72, $p = 0.000$	**
Shortenings	142	11.62%	1080	88.38%	(1, N=1222) = 11.33, $p = 0.001$	*
Clippings	8	4.68%	163	95.32%	(1, N=171) = 14.42, $p = 0.000$	**
Single-letter forms	33	7.60%	401	92.40%	(1, N=434) = 18.88, $p = 0.000$	**
Letter homophones	18	8.11%	204	91.89%	(1, N=222) = 8.40, $p = 0.004$	*
Number homophones	6	11.76%	45	88.24%	(1, N=51) = 0.43, $p = 0.510$	
Symbolic Substitution	71	21.71%	256	78.29%	(1, N=327) = 11.29, $p = 0.001$	*
Conjunctions / disjunctions	57	28.36%	144	71.64%	(1, N=201) = 27.76, $p = 0.000$	**
Punctuation omission	87	11.46%	672	88.54%	(1, N=759) = 7.70, $p = 0.006$	*
Non-standard use of lowercase	341	17.41%	1618	82.59%	(1, N=1959) = 8.40, $p = 0.004$	*
State abbreviations	4	10.00%	36	90.00%	(1, N=40) = 0.80, $p = 0.371$	
Onomatopoeic expression	74	13.73%	465	86.27%	(1, N=539) = 0.75, $p = 0.386$	
Phonetic respellings	26	4.33%	575	95.67%	(1, N=601) = 54.17, $p = 0.000$	**
Offsetting punctuation	12	3.42%	339	96.58%	(1, N=351) = 37.21, $p = 0.000$	**
All caps	69	6.73%	956	93.27%	(1, N=1025) = 55.63, $p = 0.000$	**
Letter duplication	23	9.24%	226	90.76%	(1, N=249) = 6.61, $p = 0.010$	
Punctuation duplication	202	21.04%	758	78.96%	(1, N=960) = 26.80, $p = 0.000$	**
Spelling aloud	3	60.00%	2	40.00%	--	
Emoticons	132	30.63%	299	69.37%	(1, N=431) = 81.57, $p = 0.000$	**
Emotes	19	8.15%	214	91.85%	(1, N=233) = 8.70, $p = 0.003$	*
Pointing	0	0.00%	18	100.00%	--	
Pictograms	0	0.00%	1	100.00%	--	
Misspellings / typos	166	26.48%	461	73.52%	(1, N=627) = 63.80, $p = 0.000$	**
Repairs	1	4.35%	22	95.65%	--	
Addressivity	5	100.00%	0	0.00%	--	

Feature	Serious		Recreational		χ^2	Sig
	n	%	n	%		
Reduplication	2	9.09%	20	90.91%	--	
Affixation / combining forms	26	20.47%	101	79.53%	(1, N=127) = 2.90, $p = 0.088$	
Compounds / space omission	121	17.16%	584	82.84%	(1, N=705) = 2.43, $p = 0.119$	
Blends	0	0.00%	1	100.00%	--	
Conversion	2	1.89%	104	98.11%	(1, N=106) = 14.39, $p = 0.000$	**
Formatting workarounds	9	42.86%	12	57.14%	--	
Total	2072	15.06%	11682	84.94%		
** $p < .001$						
* $p < .01$						

If one assumes that in recreational discussions, interlocutors are more likely to produce playful or creative variations, these results would support that supposition. Shortenings, clippings, single-letter forms, letter homophones, punctuation omission, phonetic respellings, offsetting punctuation, all caps, letter duplication, emotes, and conversion comprise the list of significant features that appear in greater proportions in recreational contexts. The majority of these involve more play with typography and orthography (and in the case of conversion, play with morphology) than the features that were higher in proportion in serious contexts. Shortenings—and to a lesser extent, all caps—are probably the most common outside of online communication.

Acronyms, symbolic substitution, conjunctions/disjunctions, non-standard use of lowercase, punctuation duplication, emoticons, and misspellings/typos have significant chi-square values and are higher in proportion in serious discussions. Acronyms and conjunctions/disjunctions are seen frequently in general/standard English writing, so their inclusion in this list is unsurprising; there may be nothing novel or innovative in their use. Emoticons and non-standard use of lowercase, which will be discussed more fully in the Discussion section, are rather commonplace online, and one could make a case that they are a matter of convention.

The higher proportion of misspellings and typos in serious conversations is interesting; because although one might assume that interlocutors who participate in serious discussions would be more concerned with satisfying norms and strictures for correctness (as appears to be the case in other forms of serious writing, such as scholarly writing), interlocutors represented in this corpus are not more likely to observe spelling and grammatical dictums in an effort to comply with the seriousness of the tone.

Table 31 compares non-recreational contexts (i.e., Serious and Ambiguous) with recreational/leisure-oriented contexts. The Mixed category was removed from this comparison.

Table 31: Comparison of features by non-recreational (serious and ambiguous) and recreational/leisure-oriented purposes.

Feature	Non-recreational		Recreational		χ^2	Sig
	n	%	n	%		
Acronyms / initialisms	555	22.56%	1905	77.44%	(1, N=2460) = 33.47, $p = 0.000$	**
Shortenings	180	14.29%	1080	85.71%	(1, N=1260) = 12.20, $p = 0.000$	**
Clippings	9	5.23%	163	94.77%	(1, N=172) = 19.15, $p = 0.000$	**
Single-letter forms	43	9.68%	401	90.32%	(1, N=444) = 521.10, $p = 0.000$	**
Letter homophones	19	8.52%	204	91.48%	(1, N=223) = 13.74, $p = 0.000$	**
Number homophones	6	11.76%	45	88.24%	(1, N=51) = 1.37, $p = 0.242$	
Symbolic Substitution	86	25.15%	256	74.85%	(1, N=342) = 11.56, $p = 0.001$	*
Conjunctions / disjunctions	66	31.43%	144	68.57%	(1, N=210) = 25.30, $p = 0.000$	**
Punctuation omission	103	13.29%	672	86.71%	(1, N=775) = 11.97, $p = 0.001$	*
Non-standard use of lowercase	395	19.62%	1618	80.38%	(1, N=2013) = 3.26, $p = 0.071$	
State abbreviations	4	10.00%	36	90.00%	(1, N=40) = 1.76, $p = 0.185$	
Onomatopoeic expression	90	16.22%	465	83.78%	(1, N=555) = 1.29, $p = 0.256$	
Phonetic respellings	38	6.20%	575	93.80%	(1, N=613) = 58.37, $p = 0.000$	**
Offsetting punctuation	16	4.51%	339	95.49%	(1, N=355) = 44.12, $p = 0.000$	**
All caps	85	8.17%	956	91.83%	(1, N=1041) = 69.01, $p = 0.000$	**
Letter duplication	26	10.32%	226	89.68%	(1, N=252) = 10.24, $p = 0.001$	*
Punctuation duplication	249	24.73%	758	75.27%	(1, N=1007) = 30.11, $p = 0.000$	**
Spelling aloud	3	60.00%	2	40.00%	--	
Emoticons	162	35.14%	299	64.86%	(1, N=461) = 90.70, $p = 0.000$	**

Feature	Non-recreational		Recreational		χ^2	Sig
	n	%	n	%		
Emotes	19	8.15%	214	91.85%	(1, N=233) = 15.48, $p = 0.000$	**
Pointing	0	0.00%	18	100.00%	--	
Pictograms	0	0.00%	1	100.00%	--	
Misspellings / typos	189	29.08%	461	70.92%	(1, N=650) = 53.16, $p = 0.000$	**
Repairs	1	4.35%	22	95.65%	--	
Addressivity	5	100.00%	0	0.00%	--	
Reduplication	2	9.09%	20	90.91%	--	
Affixation / combining forms	29	22.31%	101	77.69%	(1, N=130) = 1.57, $p = 0.210$	
Compounds / space omission	177	23.26%	584	76.74%	(1, N=761) = 13.82, $p = 0.000$	**
Blends	0	0.00%	1	100.00%	--	
Conversion	2	1.89%	104	98.11%	(1, N=106) = 18.767, $p = 0.000$	**
Formatting workarounds	18	60.00%	12	40.00%	(1, N=30) = 35.62, $p = 0.000$	**
Total	2577	18.07%	11682	81.93%		
** $p < .001$						
* $p < .01$						

With the exception of compounds/space omission and formatting workarounds, all features that are significant in Table 31 were also significant in Table 30, and their prevalence in either recreational/leisure-oriented contexts or serious (in this case, non-recreational) is the same. Compounds/space omission appeared in higher proportions in non-recreational contexts, and formatting workarounds appeared more often in recreational contexts.

Conclusion

These findings show that cyberlanguage feature use differs based on media and genre factors. Research question 1 asked what cyberlanguage features are common across media and genre situations. “Common” could be defined as those features that appear frequently. Table 19 helps to address this question by showing each feature’s frequency and its

percentage out of all cyberlanguage tokens in the corpus. Using this definition of “common,” the features at the top of this table would appear to be the most common.

Another approach to this question would be to define “common” as those features that did not yield significant chi-square values in all five media, the three core topics (Gaming, Technology/Gaming Technology, and Other), and the two core purposes (Serious, Recreational), and would thus be interpreted as demonstrating some homogeneity across media, topics, and purposes. Table 20 comparing all five media, Table 28 comparing the three core topics, and Table 30 comparing the two core purposes provide data for this approach. Table 32 below shows those features (marked with an x) that did not produce significant chi-square values for each of these three main comparisons.

Table 32: Features that are common to the five media, to the three core topics, and the two core purposes; “x” signifies features with insignificant chi-square values.

Feature	Five Media (Table 20)	Three Topics (Table 28)	Two Purposes (Table 30)
Number homophones			x
State abbreviations			x
Onomatopoeic expression		x	x
Letter duplication			x
Misspellings / typos		x	
Affixation / combining forms		x	x
Compounds / space omission		x	X

No features were insignificant in Table 20, which suggests that features are not homogeneously distributed across media. Feature frequencies do vary based on the medium used for communication. This further suggests that medium and thus media factors (e.g., synchronicity) are good discriminators of online language variation. Crystal (2006, p. 271) asks if online language is a “homogeneous linguistic medium” or if it is a “collection of

distinct dialects.” He suspects that the latter is the case; and where medium is concerned, these results lend support to his supposition.

Table 32 shows that some features were homogeneously distributed across topics (e.g., onomatopoeic expression) and purposes (e.g., number homophones). These features did not have significant chi-square values in Tables 28 and 30. For example, onomatopoeic expression is common across all topics and all purposes; but misspellings and typos are common only across all topics. Thus, genre factors may be better for uncovering homogeneity in feature use than media factors are. These sets of homogeneous features are small in size (i.e., no more than six features are common across purposes and no more than four across topics). This would suggest that, where most features are concerned, variation is more likely to occur when medium, topic, and purpose differ. This leads to research question 2.

Research question 2 asked what cyberlanguage features differ between media and genre situations and how they differ. Table 33 below provides a high-level answer to this question by summarizing the differences found in Tables 21 through 26, 28, and 30. Marks are placed in cells where features were highest in proportion and had significant chi-square values. For example, single-letter forms were higher in proportion in asynchronous media than synchronous media, and the chi-square value for this comparison was significant.

Table 33: How features vary in different online communication situations; “x” signifies higher than expected frequency for statistically significant comparisons.

Feature	Synchronicity	Asynchronicity	1:1	1:N	N:N	Extended Persistence	Limited Persistence	Greater Anonymity	Lesser Anonymity	Limited Message Length	Unlimited Message Length	Easy Composition	Difficult Composition	Partially Difficult Composition	Easy Viewing	Difficult Viewing	Gaming	Technology / Gaming Tech	Other	Serious	Recreational
Acronyms / initialisms			x								x				x			x		x	
Shortenings			x														x				x
Clippings			x							x						x			x		x
Single-letter forms		x	x			x			x	x						x			x		x
Letter homophones		x	x			x			x	x						x			x		x
Number homophones		x	x			x			x	x						x			x		x
Symbolic substitution		x				x		x			x							x		x	
Conjunctions / disjunctions		x				x					x							x		x	
Punctuation omission	x				x		x	x		x											x
Non-standard use of lowercase	x		x				x	x		x				x							x
State abbreviations	x						x			x											
Onomatopoeic expression	x				x		x	x													
Phonetic respellings	x				x		x			x											x
Offsetting punctuation		x		x		x			x		x										x
All caps	x				x		x	x		x											x
Letter duplication	x				x		x	x		x											x
Punctuation duplication				x				x			x										x

Feature	Synchronicity	Asynchronicity	1:1	1:N	N:N	Extended Persistence	Limited Persistence	Greater Anonymity	Lesser Anonymity	Limited Message Length	Unlimited Message Length	Easy Composition	Difficult Composition	Partially Difficult Composition	Easy Viewing	Difficult Viewing	Gaming	Technology / Gaming Tech	Other	Serious	Recreational
Emoticons		x				x			x		x		x			x				x	
Emotes	x				x		x			x				x							
Misspellings / typos			x									x									
Repairs																					
Affixation / combining forms												x									
Compounds / space omission		x		x		x			x		x										
Conversion	x				x		x										x				x
Formatting workarounds		x																			

Tables 20 through 31 suggest that there may be some validity to the idea that the constraints and affordances of cybermedia influence communication and lead to linguistic variation. The researcher would argue, however, that these results do not suggest a hard-line, technologically-deterministic perspective where interlocutors have no or very little sense of agency when producing language. Herring, Stein, and Virtanen (2013, p. 7) explain that this view—where “user behavior is a result of the physical conditions of production and reception of the medium”—originated with the application of Daft and Lengel’s (1984) theories to online communication. Instead, a more flexible interpretation would be that interlocutors are more likely to use some features over others depending on the medium, because some features may be more effective in one medium over others for achieving certain goals, such as establishing social presence and conveying social meaning. For example, Table 21 shows higher frequencies of emotes in synchronous media and this difference is significant. Virtanen (2013) believes emotes are able to accomplish social action, so perhaps synchronous situations lend themselves better to or call for performativity in order to better establish sociality. This interpretation brings Walther’s (1992) and Walther, Loh, and Granka’s (2005) social information processing theory to bear on these results. This theory—as discussed in the Literature Review—rejects applications of Daft and Lengel’s (1984, 1986) ideas to online communication and instead suggests that interlocutors modify their communication by creating textual surrogates for missing face-to-face cues to make conversation more personal and to convey relational information.

The Discussion section that follows focuses on specific features and provides a deeper discussion of how they vary in different situations. Examples of terms will be

provided to illustrate points made about this variation as well as to support any assertions made about creativity in answer to research question 3.

Discussion

Major Findings

From these results three overarching findings become clear. (1) Contrary to current thinking about “technological determinism,” technology does indeed have some association with—some possible influence over—language production. (2) New terms are being created all the time online and this suggests rapid language change. However, certain cyberlanguage terms and features are quite ordinary and conventional. (3) Cyberlanguage assumes a small portion of the language used online, so fears about cyberlanguage signaling the demise of “proper” English can be allayed.

Technological determinism

Early analysis of cyberlanguage focused on identification of features in different media for the purpose of sifting out any influence the medium might exert on communication behavior (Herring, Stein, & Virtanen, 2013). These analyses were influenced by theories about media richness (discussed in the Literature Review of this dissertation) by Short, Williams, and Christie (1976), Daft and Lengel (1986), Sproull and Kiesler (1986), and Carlson and Zmud (1999). In various ways, these theories suggest an association between media and communication behavior, which is referred to as the technologically deterministic perspective of online language where “user behavior is a result of the physical conditions of production and reception of the medium” (Herring, Stein, & Virtanen, 2013, p. 7). “In this

view, language and language usage are shaped by the constraints and affordances of the medium” (Herring, Stein, & Virtanen, 2013, p. 7). Herring, Stein, and Virtanen (2013, p. 7) claim that this view is “intuitively true to some extent,” and that claims about the association between media and communication behavior have been made with varying degrees of strength. These claims are not as authoritative as one might like because researchers have typically analyzed one medium in isolation. Without comparing language across media and communication situation, at best, one can only make assertions or suppositions about the effects media might have on communication behavior. For example, identifying high frequency of abbreviations in chat *may* be evidence that a short message length, many-to-many participant scale, and synchronicity lead to more abbreviations; but without comparing these frequencies with those found in longer messages, one-to-one and one-to-many participant scales, and asynchronous media, such claims cannot be assured. The goal of this dissertation study was to test such claims by comparing language across multiple media and communication situations. The results show that there is, in fact, some association between language production and media, media characteristics, and genre factors. So early theorists were indeed “on to something”—technology does seem to exert some influence over communication behavior.

But how much influence do media and their characteristics exert over communication behavior, and how much control do interlocutors have over precision in message meaning? Sproull and Kiesler’s (1986) views suggest that interlocutors may have little autonomy and flexibility when attempting to convey certain nuances of meaning, i.e., that interlocutors are almost puppets of the medium. This extreme view is not borne out of these results. Walther’s (1992) social information processing theory is more appropriate for this research. It suggests

that interlocutors crave social contact and to achieve this, they will exploit orthography and typography to communicate on a more intimate level. So although a medium may impose certain constraints on communication, interlocutors refuse to be hampered by them and instead exploit them to suit their communication purposes. This exploitation of the medium is evidence of creativity—the ability to solve a problem by putting elements (in this case, orthography and typography) together in new ways.

For example, cyberlanguage features can be used to make the conversation seem more like speech so that interlocutors can bridge the physical distance between them when communicating online. In particular, the introduction of surrogate face-to-face cues, such as phonetic respellings and emotes, help to achieve this goal. Interlocutors rise to the challenge of a seemingly lean medium and creatively convey social and personal meanings through “fingered speech” (Gross, 2013).

Cyberlanguage also affords other communication opportunities that are not possible in face-to-face speech. Novel forms of language play, used to build rapport, rely on the ability to see written text in action, such as the use of emotes discussed in more detail in the Emotes section that follows. Certain forms of offsetting, particularly the decoration of usernames with duplicated punctuation, also require one to see written text. These name decorations are a unique way for an interlocutor to let others conversants know s/he is excited to see them. They allow interlocutors a wider range of choices to convey their excitement, beyond how one typically conveys happiness at seeing someone in face-to-face situations (e.g., a smile or wave), as in this creative example »©« *''''*»@}~©«*Hey Hey Howdy Angie*»©~{@« *''''*»©«.

Dresner and Herring (2010) explain that not only do emoticons allow interlocutors the ability to convey facial expression and emotion, they also have illocutionary force and enable interlocutors to clarify intentions. Many emoticons used in this corpus—specifically those that were used to punctuate an utterance (see the Emoticons section below)—seemed to be used to ensure that message meaning was interpreted in a particular way.

Emoticons, like emotes and offsetting, afford interlocutors greater control over self-presentation. In face-to-face contexts, prosodic and proxemics cues are most often conveyed reflexively, involuntarily. A smile on one's face often happens naturally as a result of some positive emotional response to the conversation. Surrogate face-to-face cues online, however, are purposively selected and “coded on to the text” (Ling, 2005, p. 347). This suggests that many surrogates carry additional semantic weight beyond conveying bare bones prosodic or proxemic information. For example, intentionally choosing to elongate sounds by duplicating letters or punctuation, as in *noooooooooo!!!*, helps the interlocutor to convey emphasis and strength of emotion. *No* is a less intense reaction than *noooooooooo!!!*

This ability to exercise control over face-to-face information and to make explicit the tacit workings of one's heart and mind, afforded by the written qualities of the language, is unique to cyberlanguage⁴⁴ and is yet more evidence that interlocutors are not slaves to the medium, and that communication in online media is not necessarily deficient in comparison to more traditional forms of communication. Online communication is “not so much impoverished relative to speech and writing as different in nature from them” (Herring, Stein, & Virtanen, 2013, p. 8). It is part compensation for situational complications, and part

⁴⁴ This does not mean that writers could not purposively use linguistic features to accomplish the same ends in more traditional writing. It is merely stating that this specific kind of control of self-presentation where nonverbal information is “coded on to text”—in particular, the use of emotes, emoticons, phonetic respellings, etc.—seems to have originated in online conversation and appears to be predominantly used in online, conversational media.

celebration of a creative outlet. It affords communication opportunities that are not possible or not often pursued in traditional speech and writing, and interlocutors exercise creativity and flexibility in their communication to achieve social and personal meanings.

Ordinariness and conventionality

Despite the high frequency of hapax legomena⁴⁵ introduced over the nine years of conversations collected for this corpus that suggests rapid language change, many terms found in the corpus are quite ordinary. *LOL* (*laughing out loud*), onomatopoeic laughter (e.g., *haha*), and the smiley emoticon (:)) are a few examples. They were used in high frequencies and appear to be common means of expression. Features such as non-standard use of lowercase and acronyms were also found in high frequencies. Mostly-lowercase utterances are not something one might expect in more traditional forms of writing, and so non-standard use of lowercase may be an accepted convention of online writing. Acronyms, however, are prevalent in other forms of communication; so they may be characteristic of communication in general. What may be special about acronyms in online communication is their high frequency in comparison to that in standard forms of writing—something this study did not assess but would be prudent to pursue in future work.

Some cyberlanguage terms find their way into mainstream speech and writing. *LOL* and *OMG* (*oh my god*) are two examples of terms that have been added to the Oxford English Dictionary (Morgan, 2011). Crystal (2008b, p. 27) comments that many cyberlanguage features, such as acronyms, are “centuries old.” Novelty exists when these features are used to create new terms, but once a term is added to the Oxford English Dictionary and/or is used

⁴⁵ A *hapax legomenon* is a “word which occurs only once in a text, author, or extant corpus of a language” (Crystal, 2008a, p. 224).

in high frequencies across a corpus, it may have become a matter of convention, outliving its original innovation.

If there is a core set of vocabulary or a core set of features that comprise some sort of cohesive cyberlanguage register, then these examples may constitute that register. However, claiming any form of cohesive register is probably too strong of an assertion given the results in Tables 32 (features that are common to the five media, three core topics, and two core purposes) and 33 (feature variation across comparisons). It may be best to conceive of these as cyberlanguage staples instead.

“Proper” English

Popular press accounts of the fear that cyberlanguage signals the demise of “proper” English, as reported by Crystal (2006, 2008b), may be intensified by the high frequency of hapax legomena among the 6,604 terms containing cyberlanguage features (5,211 or 78.91%), the addition of cyberlanguage terms to well-know dictionaries, and suggestions that cyberlanguage may be an emerging, cohesive register. The goals of this section do not, however, include an attempt to incite more fear. In fact, the results of this research provide a different picture of the language.

For a variety of reasons, which will not be illuminated in depth here, no speech or writing ever adheres 100% to standard rules of grammar and spelling. For every rule one can find in a particular grammar or style guide, there is yet another rule in other grammars or style guides that contradicts. Thus, any sort of “perfect” ideal is really a matter of opinion. No agreed-upon “perfect” ideal calls into question the validity of any fears about cyberlanguage signaling the demise of “proper” English. The results from this analysis—i.e.,

that of all the tokens found in the corpus, only 10.75% included cyberlanguage features—further controvert such fears. Cyberlanguage is not colonizing English. Grammar, syntax, and vocabulary remain overall intact.

Cyberlanguage is instead inserted or built into a pre-existing foundation of general/standard English, in the same way that slang and jargon are incorporated into more standard forms of communication. “A slang term rarely violates sentence structure” and instead appears to be used in a manner that complies with established patterns (Eble, 1996, p. 21; see Eble’s *bogart* example). This is also true of cyberlanguage. A case in point is the use of the term *loling*. Although the use of the *-ing* suffix may appear redundant, it ensured the sentence in which it appeared flowed more smoothly and read more like a standard utterance (*i’m kinda loling at your right now*⁴⁶).

Despite any personal biases against jargon, slang, or cyberlanguage, these additions to general/standard English do serve important functions. Jargon helps a community of professionals or serious hobbyists talk more efficiently about their work. Slang helps to “reinforce social identity or cohesiveness within a group” (Eble, 1996, p. 11), a trait shared with cyberlanguage. Additionally, cyberlanguage helps interlocutors negotiate the constraints of cybermedia and exercise greater control over message meaning and self-presentation.

Adams (2008, p. 8) claims that when interlocutors use slang, they demonstrate “‘linguistic competence,’ that is, the innate human capacity to acquire and use language” (Adams, 2008, p. 8). Crystal (2008b) makes similar assertions about cyberlanguage, in his discussion of vowel and consonant contraction (referenced in more detail in the Conclusion to this Discussion). The way in which interlocutors reshape English to suit the communication situation does not necessarily mean that they have poor understanding of

⁴⁶ The *your* is a typo for *you*.

English. Instead, it demonstrates their ability to exploit rules of grammar and spelling to more effectively communicate given the constraints of the medium. Perhaps cyberlanguage is best thought of a specialized vocabulary intended for a specific type of setting and group of interlocutors (i.e., online conversational media and online interlocutors), and bearing qualities similar to slang and jargon such as its ability to add to general/standard English vocabulary.

Conclusion

As Crystal (2006, p. 271) suggested, cyberlanguage is not a singular, distinct language variety (or register) consistent across media. Most features vary based on medium, media characteristic, and genre factor. Thus, technology, topic, and purpose do seem to influence communication. Many new terms are created frequently and features are used in new ways to create them, but some terms are quite ordinary and some features may be used in a conventional manner. Tokens that include cyberlanguage features were much fewer in number in the corpus than general/standard English tokens. Thus, fears about cyberlanguage signaling the demise of English can be allayed. Cyberlanguage, instead, adds to general/standard English in the same way that slang and jargon do. It helps interlocutors compensate for missing face-to-face cues and to overcome other constraints of cybermedia so as to communicate on a more rich and personal level.

Specific Features

The list of results shown in Tables 20 through 33 is extensive. Only a small selection of features will be discussed in this section. Others may be discussed in future publications. Features that are discussed at some length by other researchers and for which there were noteworthy or interesting examples in the corpus are discussed in this section. These include acronyms, phonetic respellings, emoticons, and emotes (including those that are examples of pointing). Features that appeared in relatively generous numbers in the corpus and for which only passing comments are made in other research are also discussed. These include symbolic substitution and onomatopoeic expression. Some features are discussed because examples in the corpus showed uses or construction methods that have not been identified in other studies or have been discussed infrequently. These include offsetting, repairs, addressivity, and compounds/space omission. Finally, a short discussion of lowercase and its contrast, all caps, appears because lowercase proper nouns and lowercase *I* were one of the highest frequency features.

Acronyms / Initialisms

Acronyms and initialisms are the highest frequency feature in the corpus (see Table 19). They appeared 2,776 times, and although they account for over 15% of the cyberlanguage features, they are infrequent in the corpus as a whole (2.03% of all 136,529 tokens). Other researchers—such as Baron (2008), Bieswanger (2007), Lewin and Donner (2002), and Ling (2005)—have found few instances of acronyms throughout their corpora, some finding even smaller percentages. For example, Bieswanger found six acronym tokens out of the 1,120 he collected (0.54%).

In this corpus, acronyms/initialisms were highest in proportion in email (24.86%), followed by chat (16.04%), forums (15.89%), IM (9.08%), and SMS (8.25%).⁴⁷ They also appear in higher proportion in these situations: 1:N (followed by N:N), limited message length, easy composition, easy viewing, Technology (including gaming technology) topics, and Serious purposes. The differences in their distribution were not significant for the synchronicity, message persistence, and anonymity comparisons. Thus, when considering the need for speed and brevity that researchers have assumed motivates the use of abbreviations, these results suggest that acronyms/initialisms are more likely to be used if there is little room to write and if there is more competition for the floor; but the degree of ephemerality of messages or whether interlocutors converse in real-time has no effect on acronym production.

Acronyms/initialisms, being higher in proportion in Serious, Technology discussions, may be the stuff of jargon. Jargon is “the vocabulary used in carrying out a trade or profession or in pursuing an interest or hobby” (Eble, 1996, p. 19). It is used for serious purposes and serious play (Adams, 2009). Technology professionals and enthusiasts use acronyms to get work done, and serious gamers—like those described by Yee (2006)—use acronyms to conduct the business of gameplay. Thus it is no surprise that acronyms/initialisms were second highest in proportion in discussions focused on Gaming topics. Acronyms/initialisms may be normal vocabulary elements of technology and gaming registers.

⁴⁷ Throughout the discussion that follows, proportions given are the number of feature counts for a particular feature (in this case, acronyms) out of all cyberlanguage feature counts in a specific category (e.g., all feature counts for email or chat or forums, etc.).

As one might expect, many acronyms were used to refer to things—nouns and named entities—such as *nyc* (*New York City*), *WSG* (*Warsong Gulch*, a WoW battleground), *AWS* (*Amazon Web Services*), and *BC* (*Burning Crusade*, an expansion of the original WoW game). Many of these named-entity acronyms comprise the vocabulary—or jargon—of technology professionals and serious gamers.

Many other acronyms, however, were used to shorten ordinary, colloquial phrases including exclamations, greetings, closings, well wishes, and politeness markers. *Omg* (*oh my god*), *omfg* (*oh my fucking god*), *wth* (*what the hell*), and *wtf* (*what the fuck*) are examples of exclamations.

Cho (2010) claims that greetings and closings are phatic. In this corpus, greetings include *gm* (*good morning*) and *wb* (*welcome back*). There were several acronyms that served as “leave-taking formulas,” to use Cho’s (2010) terminology. They enabled interlocutors to say goodbye, put the conversation on hold in some way, or let the leaving person know a swift return is desired. Examples include *brb* (*be right back*), *bbiam* (*be back in a minute*), *bbs* (*be back soon*), *gtg* (*got to go*), *hb* (*hurry back*), *HBN* (*hurry back now*), *tc* and *t/c* (*take care*), *h/o* (*hang on*), and *ttyl* (*talk to you later*). Some of these also function as well wishes, such as *tc* and *t/c*. Other well wishes include *gg* (*good game*), *gj* (*good job*), *hagd* (*have a good day*), *hbd* (*happy birthday*), and *wtg* (*way to go*). *Imy* (*I miss you*) shows affection.

The most frequent sets of acronyms were those that are intended to act as surrogate proxemic cues, specifically those that convey laughter. The most frequent of these was *lol* (*laughing out loud*), which appeared in a variety of forms including lowercase (*lol*), uppercase (*LOL*), camel case (*LoL*), and with a zero in place of the letter *o* (*L0L*), etc. There were 65 different versions (or types) of *lol*, yielding 597 tokens (21.51% of all acronyms).

Crystal (2008b) explains that very few acronyms are used repeatedly, but *lol* is one of them. *Lol* appeared as 0.41% of the total words in Tagliamonte and Denis' (2008) corpus, a figure almost equal to that found in this corpus. North (2007) does not consider *lol* to be creative because it is so commonly used; at one time it was creative, but now it is a matter of convention in some circles. Other laughter acronyms include *lmao* (*laughing my ass off*), *lmfao* (*laughing my fucking ass off*), and *rofl* (*rolling on the floor laughing*). Although it is possible (or even likely) that when interlocutors use these terms they are not actually physically laughing or rolling on the floor in fits of giggles, these terms are probably still meant to create solidarity—acting almost as minimal responses—by confirming that the joke was heard, understood, and appreciated. An inward laugh and outward grin communicated as *lol* when participants cannot see each other's faces may go a long way in establishing rapport.

Although there are more acronyms conveying friendship and care, there were seven types that could be viewed as hostile: *DGAF* (*don't give a fuck*), *fo* (*fuck off*), *FU* (*fuck you*), *GTFO* (*get the fuck off*), *RTFM* (*read the fucking manual*), *rtfq* (*read the fucking quest*), and *stfu* (*shut the fuck up*). These accounted for eight tokens. Silva (2010, p. 270) explains that insiders may “create mechanisms to shut other users out.” The interlocutors who used these acronyms may be attempting to do this based on the context in which they were used.

Cherny (1999) claims that abbreviations are culturally conditioned. Werry (1996) explains that “certain forms of abbreviation” emerge “that are native” to a particular community. In this corpus, *LF* (*looking for*) and its variants such as *lfw* (*looking for work*), *lfg* (*looking for group*), *LF3M* (*looking for three more*), *LF2Tanks* (*looking for two tanks*) are WoW-specific. *Dps* (*damage per second*), *omw* (*on my way*), *brt* (*be right there*), *wtb* (*want*

to buy), and *wts* (*want to sell*) are others. These acronyms tell us something about the game world; it has economic structure that is built by buying and selling goods and services (*wts*, *wtb*, *lfw*). Players frequently search for other players with whom to play (*LF*, *lfg*, *LF3M*, *LF2Tanks*), thus collaboration is an important component of the game, and players assist each other with game tasks and provide feedback to their comrades to let them know to expect their help (*omw*, *brt*). In a virtual world where one participates through avatars, friends can *be there* with you. Baron (2003, p. 95) claims that abbreviations such as acronyms may “indicate one’s membership among network cognoscenti.” Being able to speak the language of a community affords one insider status (cf. Lave and Wenger, 1991).

Novel or out-of-the-ordinary acronyms that provide evidence for linguistic creativity include *loladins* and *loling*. In WoW, players may choose characters of different classes, such as mages, warriors, druids, and paladins. The term *loladins* attaches the suffix *-adins* from the class *paladin* to the acronym *lol*, to create a term that jokingly suggests there is a type of character (or, more likely, player) who is prone to laughing out loud. Although the *-ing* is unnecessary because it is packed into the acronym *lol*, the interlocutor who used *loling* added the *-ing* to make his/her comment more clear: “*i’m kinda loling at your right now*” (with the *r* at the end of *you* being a typo). If s/he had not included the *-ing*, the utterance would have been awkward (*i’m kinda lol at you*), but by unpacking the *-ing* in this way, the interlocutor creatively reshapes the acronym to fit the context.

Symbolic Substitution

There were 256 symbolic substitution types and 417 tokens. They appeared in greater proportions in forums (5.23%), followed by IM (2.71%), email (2.50%), chat (1.63%), and

SMS (0.89%). They also appeared in greater proportions in these situations: asynchronous, extended persistence, greater opportunity to maintain anonymity, unlimited message lengths, easy composition and viewing, Technology topics, and Serious purposes.

Symbolic substitution is a keystroke-saving word-creation strategy (Ferrara et al., 1991; Hård af Segerstad, 2002). Yet they were not higher in proportion in the typical time-sensitive, space-saving scenarios such as synchronous media, media with limited message lengths, or N:N situations. (In fact, differences in participant scale were not significant.) So although symbols may be used as a way to save time and keystrokes, these results suggest that there are other reasons for using them.

Their higher proportion in Technology topics is telling. Cherny (1999, p. 92) points out that “use of abbreviations and shortenings in a register is very much culturally conditioned.” Those who work in information technology or who think about it enough to motivate them to discuss it online—especially programmers—are often conditioned to think in programming terms, which use non-alphabetical characters in symbolic ways. Cherny (1999) provides several examples of terms that draw on programming lingo. The language used in the MOO she studied was driven, in large part, by the people who programmed the MOO. MOO administrators thought about language in terms of functionality. If a term or concept was popular or particularly interesting to the MOO administrators, they automated it by creating a command that would enable interlocutors to quickly and effortlessly include it in their communication. This is how many emotes came to be used. For example, one of the administrators wanted the ability to eye someone warily, so he programmed the *>eye* command into the MOO so that he and other users could give someone the eye quickly and easily. Typing *>eye* is faster than typing *Laura eyes Stephanie warily*, which is what was

output to the screen upon using the command. Programming and symbolism was an integral part of the communication in Cherny's MOO. The same may be said here when reflecting on the higher proportion of symbolic substitution in technology topics. Information technologists may simply be equally comfortable working with symbols as with words, and so although they do not need to be brief, they are naturally more inclined to write with symbols.

As was discussed in the Compositional Ease section of the Findings, typing punctuation on mobile devices may require more effort and be potentially frustrating because of the interface and input device's inflexibility with regard to punctuation. So the higher proportion of symbolic substitution in media with easy composition and viewing is unsurprising.

Most instances of symbolic substitution were also conjunctions and disjunctions: 159 types (62.11% of symbolic substitution types) and 241 tokens (57.79% of symbolic substitution tokens). WoW players often play several characters, each with a different username/character name. To associate all their characters so that other players may consolidate their view of these characters into one player, players may list out their character/usernames intercalated with slashes. For example, the researcher could list her own characters as follows: *Innle/Nawyn/Skygge/Shinzui/etc*. In this utterance,

the worst part about origins was combat being slow/boring/same thing which they seemed to fix in DA2

the interlocutor is explaining that the worst part of the game Dragon Age Origins was that combat was boring, slow, and routinized. This was fixed in Dragon Age 2, the second

version of the Origins game. This use of slashes is a common way to create conjunctions and disjunctions in the corpus.

Another common usage is the *a/s/l* (*age, sex, and location*) formula used in AOL and NPS chat. This is a way to introduce one's self to the chatroom, and it exposes an unspoken assumption that many chatters are there for romance. If a person's age, sex, and location fits with the qualities one is seeking in a partner,⁴⁸ a personal conversation may be initiated. Some examples include *19/f/GA* and *17/m/oh*. Some interlocutors will stretch the formula to include other descriptive information such as *21/f/bored*, *21/f/single*, *21/m/big....*, *24/m/white*, and *single/man/35/New York*. If a user logs on and does not produce this information, s/he may be asked "a/s/l?"

Punctuation is also used to signify acronymy, as in *t/c* (*take care*), *b/c* (*because*), *f/t* (*full time*), *d/c* (*disconnected*), *h/o* (*hang on*), *j/k* (*just kidding*), and *w/l* (*wireless*). Perhaps the slash signals to others that the term should be recognized as an acronym, particularly if the term has a non-acronymized sense. For example, *h/o* was used in this short utterance: "h/o k". Without the slash, it might read as "ho k" which could be interpreted as a phonetic respelling of *oh* and a single-letter form of the *-kay* part of *okay*.

Non-alphabetic characters may also be used to filter profanity. For example, *kick-a*** (*kick-ass*) and *F#\$ed* (*Fucked*)⁴⁹ are ways interlocutors disguise their curse words. This may suggest that some interlocutors wish their utterances to convey the force or emphasis associated with the use of profanity, but without overly offending sensitive ears. It could also be for cartoony effect as well. This usage of symbols along with slashes used for acronymy

⁴⁸ The term *partner* is not meant to suggest that chatters are looking for marriage or long-term relationships outside of the chat room necessarily. Some may be, while some may be looking for a romantic partnership that is a bit more fleeting and limited to the Internet.

⁴⁹ This example is not a one-to-one replacement.

often does not result in abbreviated forms. Thus, symbolic substitution may also be used for clarity and affect.

Other noteworthy examples of symbolic substitution include carets (^) used to signify agreement with a preceding comment in a chat conversation (one example in the corpus duplicated the caret to signify emphatic agreement (^^)), and a caret used as a way to signify a high five (^5). WoW players use a symbolic formula—noun x numeral—to signify a quantity of an item. Examples include *Saronitex6* and *Frozen Orb x3*. This formula is most often used when players wish to trade or sell items. Lindh (2009) also identified this formula in his analysis of WoW chat, and some of his examples use plus signs (+) or asterisks (*). The letter *x* was used to signify a kiss, the letter *o* to signify a hug. Although these are both long-standing forms of symbolic substitution, only three types were found (*xoxo*, *xo*, *xx*). Bieswanger (2007) also found a few instances (4 types) in his SMS corpus.

Lowercase and All Caps

Online, “there is a strong tendency to use lower case everywhere” (Crystal, 2006, p. 90). Crystal (2006) refers to this as a “lower-case default mentality” (p. 92). Because of this “any use of capitalization is a strongly marked form of communication” (Crystal, 2006, p. 92). Capitalization may be used for shouting or emphasis (Danet et al., 1997).

The most frequent cyberlanguage word in the corpus was a lowercase pronoun *I* (1,067 tokens, 7.27% of all 14,681 cyberlanguage tokens). Non-standard use of lowercase was the second most frequent cyberlanguage feature in the corpus (2,606 tokens, 14.28% of all cyberlanguage features). Even though analysis of non-standard lowercase was limited to proper nouns and the pronoun *I*, these figures support Crystal’s assertions.

Non-standard use of lowercase appeared in highest proportions in IM (22.44%), followed by chat (14.64%), forums (13.38%), SMS (12.84%), and email (7.07%). IM, chat, and forums are all typically accessed from a full-sized computer and keyboard; so users should encounter fewer difficulties capitalizing letters than SMS users. Thus, the size and/or difficulty of the input device are probably not motivators for neglecting the shift key. Neither does time appear to be a motivator. Non-standard lowercase appeared in greater proportions in asynchronous and 1:1 media that apply less pressure on interlocutors to communicate quickly. They were greater in proportion in limited persistence media and media that impose message length restrictions, but capitalizing letters does not require additional character space. Perhaps lowercase is used simply because it requires less effort and is an accepted way of communicating online. It is used in greater proportions in gaming conversations, so perhaps it is at least a convention in gaming communities. Non-standard lowercase also appeared in greater proportions in conversations for Serious purposes, which suggests that the convenience of lowercase outweighs any grammatical strictures one might expect to find in more serious conversations.

Typing words in all caps is done with less frequency than not capitalizing proper nouns and the pronoun *I*. These all-caps terms would tend to visually stand out in a sea of lowercase as Crystal suggests. However, it is still a high-frequency feature, the fifth-highest listed in Table 19. There were 1,229 tokens (6.74% of all cyberlanguage features) and 675 types. These were greatest in proportion in chat (9.25%), followed by forums (5.74%), IM (5.53%), email (2.61%), and SMS (1.74%). As such, all caps are used more in synchronous, N:N, limited persistence, more anonymous, limited message length, partially difficult composition, and easy viewing media. Those instances of all caps that are interpreted as

shouting may be examples of flaming, which Herring (2002) says is due to the low social accountability that pervades more anonymous situations.

Onomatopoeic Expression

Few researchers discuss onomatopoeic expression, yet in this corpus they appear almost as frequently as emoticons, which have received considerable attention. There were 327 types of onomatopoeic expression, 678 tokens. They appear most frequently in IM (4.91%), followed closely by chat (4.37%), SMS (2.54%), forums (2.51%), and email (2.34%). They appear in greater proportions in synchronous, N:N, more anonymous media that place limits on message length and enable partially difficult composition and easy message viewing. Some of these media characteristics—such as synchronicity, N:N participation, and greater opportunities for anonymity—are thought to create fertile ground for play and performance (Crystal, 2006; Danet, 2001; Danet et al., 1997; Werry, 1996). This suggests play and performance may be chief activities associated with the use of onomatopoeic expression. Cherny (1999, p. 113) believes the exclamations and interjections that comprise, in part, this dissertation's definition of onomatopoeic expression are a "type of modality play."

Most of the onomatopoeic expression in the corpus was either some form of exclamation or minimal response. Examples of exclamations and other vocalizations, which appeared in a variety of forms, often with letter duplication, include *ew*, *ugh*, *aw*, *oo*, *Yippee!!!!*, *Yay!*, *woohoo*, *wooo*, *woot*, *whoaaaaaaaa*, *oops*, *woops*, *hooray!*, *ARGH!!*, *wEEEEEEEE*, *boo*, *bah*, *eek*, *grrrrrr*, *pfffft*, *shhhhhhhhhhhhhhh*, *whew*, *yow*, and *ouch*. Many of these have a cartoony feel to them, and as such invite play and laughter.

The most frequent form was *oh* and its variants, such as *Ohhhhhhhhhhhhhhhhh* and *oooooooooooooooooooooh*. There were 26 types of *oh*, and 178 tokens. *Oh* can be used as both an exclamation and a minimal response. The second most frequent expression was spelled out laughter: *haha* or *hehe* and their variants, also forms of feedback and exclamation. There were 38 types of spelled-out laughter and 88 tokens. Between all the variants of *haha*, *hehe*, and the second most frequent laughter type in the corpus (*lol*), there is plenty of laughter going on in this corpus (685 tokens). Perhaps this laughter is another indicator of play and expressivity.

The several, more familiar types of minimal responses and their variants, which often included letter duplication, appeared in the corpus, such as *hm* (23 types, 37 tokens), *mm* (4 types, 6 tokens), *mhm* (4 types, 8 tokens), *ah* (15 types, 38 tokens), *um* (11 types, 18 tokens), *eh* (4 types, 7 tokens), and *er* or *erm* (8 types, 8 tokens). They total 122 tokens (17.99% of all onomatopoeic expression found in the corpus). It appears that in the online situations represented in the corpus, interlocutors attempt to signal their attention to the conversation, which probably helps build rapport and group cohesion.

Only a few types of onomatopoeic expression were true sound effects. These include *ding*, *CHA CHING*, *boom*, *poof*, *Puff*, *whoosh*, *achhoooo*, and various kissing sounds like *muahs* and *muahz*. Herring (2012, p. 3) points out that “non-language sounds enrich CMC in the absence of auditory cues”; however, because of the infrequent instances of such sounds, it may be assumed that of the types of aural cues that may be included in utterances, sound effects are not a popular one.

Phonetic Respellings

Phonetic respellings appeared 812 times in the corpus (4.45% of all cyberlanguage features in the corpus) and in 333 types. They were highest in proportion in SMS (6.87%), followed by chat (5.66%), forums (2.73%), IM (1.25%), and email (1.14%). According to Thurlow (2003) surrogate prosodic cues, like phonetic respellings, help to achieve a playful, informal tone that befits the relational nature of SMS. The same might be said of phonetic respellings in chat, given that they appear there almost as frequently as they do in SMS.

Phonetic respellings appeared in higher proportions in synchronous, N:N, limited persistence media with limited message lengths and challenging user interfaces. This suggests that in addition to their ability to infuse online conversation with prosody, they may also serve as a way to abbreviate messages. Hård af Segerstad (2002), Herring (2012), Thurlow (2003), and Thurlow and Poff (2013) all comment on their ability to save keystrokes, time, and effort. Many-to-many situations may also act as performative spaces, and the higher proportion of phonetic respellings in these situations suggests that they may serve performative functions. This supports North's (2007) supposition that they highlight performative aspects of interaction.

Phonetic respellings also appeared in higher proportions in conversations that discussed Other topics and were used for Recreational purposes. According to Carter (2004), informality, which is expected in recreational contexts, may reinforce the desire to simulate accents. Recreational contexts are naturally oriented toward play, so phonetic respellings, especially those that imitate accents (e.g., *underdawg*, *fixn*, *fer*), may be a way for interlocutors to play with language and create an informal, highly social atmosphere. Tagg (2009) believes they create intimacy, set an informal tone, and enable users to play with

identity. Examples in the corpus that are humorous and playful include *Eggseolent* (*excellent*), *bewbs* (*boobs*), *delinkwent* (*delinquent*), *doowidda* (*do with*), *drood* (*druid*), *egz* (*eggs*), *fawk* (*fuck*), *goshnezz* (*goshness*), *hawt* (*hot*), *lubbs* (*loves*), *muahzzzz* (the sound of kissing), *phukin* (*fucking*), *smewchies* (*smoochies*), *sowwy* (*sorry*), *moar* (*more*), *kewl* (*cool*), and *dewd* (*dude*). Most of these do not shorten the word, and so are not serving brevity functions. Instead they serve phatic functions as Tagg and others suggest. For instance, several seem to be for the purpose of being silly. In a discussion about sending nude pictures of one's self, the interlocutor who used the term *bewbs* feigns innocence by asking “*what are bewbs?*”. Exaggeration, a strategy used in many jokes, is found in “*Whaddaya wanmeta doowidda PPT file?*” (What do you want me to do with the PPT file?). Some phonetic respellings appear to be intended to show support and lighten the mood. In a chatroom discussion about a participant's depression, one interlocutor offers *smewchies* to the depressed participant, and another offers love, as in “*you know i lubbs ya honey :)*”. According to Hård af Segerstad (2002, p. 219), phonetic respellings may act as “in-group markers,” connecting interlocutors through shared knowledge and shared vocabulary. These examples support this idea; phonetic respellings signal unity and belonging. Tagg (2009, p. 145) explains that they “highlight the importance of the interpersonal over physical constraints.”

Not all phonetic respellings are as inventive, playful, or silly as the ones shown above. Some are rather ordinary. Crystal (2006) explains that some respellings are so widely used that they are almost standard. He discusses several that appear in dictionaries and literary works, such as *ya*, *wanna*, *dunno*, *gonna*, *thanx*, *luv*, *sorta*, *thru*, and *skool* (Crystal, 2008b, p. 49). All of these appear in this corpus, and *ya* for *yes* or *you* is the most frequent;

wanna is second highest in frequency, *gonna* fourth, and *dunno*, *luv*, and *thru* also appear toward the top of the list.

Offsetting Punctuation

There were 261 types of offsetting and 363 tokens. They appeared most frequently in email (8.76%), followed by chat (1.86%), forums (0.65%), IM (0.26%), SMS (0.09%). Thus, they appeared in greater proportions in asynchronous media, 1:N conversations (followed by N:N), extended persistence media, media affording less anonymity, media with little to no restrictions on message length, media with easy composition and viewing, and conversations focused on Other topics for Recreational purposes. Underscores, pound signs, equal signs, carets, slashes, and asterisks are some of the punctuation that can be used to offset a word or phrase (Crystal, 2006). More often offsetting is discussed as using asterisks (Danet, 2001; Lewin & Donner, 2002). Most of the offsetting examples in the corpus used asterisks in one way or another.

Offsetting is most often discussed as a means of emphasizing a word or phrase. However, in this corpus, they are used in other ways, some of them more frequent than for emphasis. These different uses include marking emotes, decorating names as a form of greeting, marking emoticons, and substituting for quotation marks.

Some examples of offsetting for the purpose of emphasis⁵⁰ include:

****Proofread it once.****

⁵⁰ Offset terms are marked in bold.

*It is interesting, though, that ***today*** [SINGER'S]⁵¹ recordings sell better in the UK than in the US.*

*threat is ***never*** an issue in a good raid, good DPS know how to watch Omen and react accordingly*

*Now, if I have done to you -- what was done to me then the above will be burned into ***your*** brain for the next 24 years.*

Did I not just /say/ that?

Fifty-six types (63 tokens) marked emotes (see the Emotes chapter later in this Discussion for examples). Forty interlocutor usernames were decorated with parentheses, asterisks, brackets, and other punctuation. Name decoration could be viewed as a special case of offsetting, which may fit Danet's (2001) definition of ASCII art. These are, in one sense, a form of pictogram. Some examples include:

(>(*(*(*CARM)*))**)*

(((((carm)))))))))))))

******dale******

[[[[[[[[[[[[[[[HELLO ANG]]]]]]]]]]]]]]]]]]

*»©« *`´´*»@}~©«Hey Hey Howdy Angie»©~{« *`´´*»©«*

As these examples show, many interlocutors go above and beyond ordinary or simple offsetting that uses a few asterisks on either side of the term. The last example above is

⁵¹ The opera singer's name was removed to protect the identities of the participants of that email list. In its place [SINGER'S] appears.

particularly creative and artistic. The time and care that went into its creation may reflect fond feelings toward the recipient Angie.

Many “greeting rituals are particularly routinized” (Cherny, 1999, p. 116). Decorating names was done predominantly by AOL chatters, and to a lesser extent by NPS chatters. It appears that this form of greeting is conventional in AOL chat, and somewhat conventional in NPS. AOL and NPS participants predominately use offsetting for emotes and username decoration.

Seven emoticons (30 tokens) were enclosed in brackets. All were created by AOL chatters. So it appears that in addition to decorating names with punctuation, sandwiching emoticons in brackets is another convention of the AOL community. The emoticons are as follows:

<i>[:(]</i>	Frowny
<i>[:)]</i>	Smiley
<i>[:/]</i>	Lips pursed in frustration
<i>[:D]</i>	Smiley
<i>[:P]</i>	Tongue hanging out of the mouth
<i>[:)]</i>	Winking smiley
<i>[>:o]</i>	Devil face with horns and a wide open mouth

The square brackets may be a way to frame the face—in the sense that the leftmost bracket acts as the top of the face or the edge of the forehead, and the rightmost bracket acts as the

bottom of the face or the edge of the chin. If this is so, then these examples might have been better classified as only being emoticons. However, the square brackets could also be thought of as a pictographic form of offsetting in the way that name decoration is.

The most frequent form of offsetting was the use of asterisks as substitutes for quotation marks. As such, these could have easily been classified as symbolic substitution, but they were not, simply because the dominant spirit of symbolic substitution is abbreviation. Substituting asterisks for quotation marks saves no keystrokes, time, or effort, particularly if a full keyboard is available. Furthermore, all instances were constructed by members of the opera fan email list, which suggests that this is a matter of group style. Crystal (2006, pp. 196-197) explains that “each group will have its favourite jargon, its ritualized utterances, and its idiosyncratic commands.” All of the aforementioned uses of offsetting, with the exception of emphasis, appear to comprise a set of stylistic features particular to one of the groups represented in the corpus.

Emoticons

Emoticons appeared in 3.96% of cyberlanguage features (see Table 19). There were 247 types and 723 tokens. In comparison to other cyberlanguage features, such as acronyms/initialisms, non-standard use of lowercase, and shortenings, emoticons are rather scarce. Several researchers have commented on this for one or more of the five media: Dürscheid and Frehner (2013) for email, Crystal (2008b) and Thurlow (2003) for SMS, Lewin and Donner (2002) for forums, Baron (2010) and Varnhagen et al. (2020) for IM, and Cherny (1999) for chat.

The differences in emoticon distribution in every comparison were significant. Across all five media, emoticons were highest in proportion in email (9.41%), followed by SMS (6.91%), forums (4.47%), IM (3.34%), and chat (2.19%). They are more prevalent in these situations: asynchronous media, 1:N (followed by 1:1), extended persistence, lesser opportunity for anonymity, unlimited message lengths, difficult composition and viewing, Other topics, and Serious purposes.

Dresner and Herring (2010) claim that emoticons tend to occur more often in synchronous communication, but these results do not support their assertion. Their more frequent use in asynchronous, non-many-to-many, and extended persistence media suggests that, when constructing emotions, interlocutors may require more time and less pressure to communicate quickly.

Dresner and Herring (2010, p. 261) also claim that emoticons are more often used in “informal, playful communication than in formal or task-focused CMC.” These data show greater proportions of emoticons in conversation for Serious purposes, which appears to contradict Dresner and Herring’s assertions. The transcriptionists’ email list used the highest proportion of emoticons (26.68% of all 431 cyberlanguage features were found in the transcriptionists’ communication), and all emails were focused on work tasks. The second highest proportion of emoticons appeared in the beekeepers’ communication (8.90% of the 146 cyberlanguage features were found in the beekeeper’s communication); many emails focused on beekeeping tasks and serious topics such as county ordinances related to keeping bees.

Emoticons are frequently discussed as expressions of emotion (Danet et al., 1997), and surrogate facial expressions that “add a sense of face-to-face interaction to a message”

(Yongyan, 2000, p. 32). But emoticons “extend beyond substituting for facial and gestural cues” (Dresner & Herring, 2010). They also serve a variety of other phatic and social functions (Walther & D’Addario, 2001). For example, they may signal common knowledge, and help clarify message meaning (Walther & D’Addario, 2001). Tagg (2009) believes they function as “response tokens” (similar to head nods and other minimal responses) enabling interlocutors to confirm to their fellow conversants that they are listening, and as such emoticons help make the conversation seem more speech-like. Dresner and Herring (2010) claim that they have illocutionary force and are thus used to convey intention (e.g., to signify a humble request).

Many emoticons in this corpus were used to end an utterance. Often, there was no intervening space between the final word (or in several cases, a final series of ellipses) and the ending emoticon (e.g., “*Hola amor. I had the sound turned off. Please don't forget the Oxford book:-D*”). In this way, emoticons could be viewed as serving grammatical functions, but they could also be a means to communicate a final intention or final impression they would like to leave with the receiver. The *:-D* smiley in the above example could be the interlocutor’s way of apologizing for having the sound turned off (and possibly missing an earlier call), and/or his/her way of adding another layer of politeness to the command to not forget the book. In this example—“*are you going to loan these out for us to use? I am ready to meld the wax to the wires:D*”—the interlocutor may be using the smiley to soften what might otherwise come across as impatience. By including the smiley, the interlocutor may be suggesting that although s/he would like to borrow the items because s/he is ready to use them, there is no rush and no pressure to loan them.

Cherny (1999) found few emoticons in her chat corpus, but smileys were the most common types of emoticons. Similarly, the most frequent emoticon in this corpus was :). It appeared 109 times. Hård af Segerstad (2002) and Lindh (2009) also found this exact smiley (as opposed to :D or other smileys) to be the most frequent in their corpora. Other high frequency emoticons include :-), :D, :P, ;), and =). Bieswanger (2013) claims emoticons are “characteristic” of online communication. To some, they have become conventional (Dresner & Herring, 2010; North, 2007). Walther and D’Addario (2001) think they are overused, and because of this North (2007) thinks very few are creative. Some emoticons found in this corpus may be less familiar and are possibly more innovative or have been used in more unusual ways. These include:

<3	Kissy lips from a profile view or, according to Wutiolarn and Attaprechakul (2012), a heart
<33 :-*:-* :-x:-x	Three different ways to give multiple kisses
>.>	Looking off in another direction
:3	A cat’s face
<(:-)	A smiley face wearing a hat
:S	A snaky, confused mouth
- _ - ‘‘	The two single quotes represent sweat
:^0	Pointy nose and wide open mouth
>:o >:-)	Two faces with horns
<i>sweeth;-)eart</i>	A winking smile inserted in the middle of a word

Some interlocutors also emphasize their emoticons by duplicating some of the punctuation. For example, :-((for signifying feeling really sad or :() for feeling really happy.

Emotes

There were 131 types of emotes and 275 tokens. They are most often discussed in studies of chat, and in this corpus, they appeared more frequently in chat (2.07%). They appear next most frequently in email (1.63%), followed by SMS (1.56%), forums (0.33%), and IM (0.21%). As a result, emotes appear in greater proportions in synchronous media (but not by much in comparison to asynchronous), N:N (similarly, not much more than in 1:N), limited persistence media, media with limitations on message length, partially difficult composition situations, Other topics, and Recreational contexts. Their higher proportion in synchronous, N:N, limited persistence, and limited message length media suggest that some emotes may be serving an abbreviation function, such as when one-word emotes are used (e.g., *grins*). This supports Cherny's (1999) assertion that they may be used to reduce typing.

Emotes are often enclosed in punctuation and so were also classified as having offsetting punctuation. Sometimes they are enclosed in asterisks (Cherny, 1999; Werry, 1996; Wutiolarn & Attaprechakul, 2012); sometimes in angle brackets (Cherny, 1999; Herring, 2012; North, 2007); and they may also be found enclosed in parentheses (Wilkins, 1991). A review of full emotes (i.e., the entire emoted phrase, if the emote contained more than a single word) showed that most emotes in the corpus were enclosed in asterisks (23 phrases or words, 28 instances). Parentheses (e.g., *(does his best William Shatner bashing Trekkies imitation)*), brackets (e.g., <waving>), slashes (e.g., /ducks), and colons (e.g., :

hands lauren a rose:) were also used to mark an emoted word or phrase. These later forms of punctuation were used in roughly equal proportions—all appeared in four phrases or words, and occurred in four instances, except parentheses which occurred in five instances. Dashes (e.g., *-hugs tight-*) and periods (e.g., *.hugs tammy.*) were also used but were the least frequent type of punctuation employed. The majority of emotes were not marked with punctuation.

Fifteen different types of pointing arrows were used (17 instances). Most of these are intended as a way to indicate one's state of being, and as such qualify them as "exposition" emotes (Cherny, 1999). For example, <<<<<<<*confused*, <*is a smoker*, and <*sorry dont drink*⁵² indicate one's emotional state or personal characteristics.

Most emotes appear in third person singular present tense (Cherny, 1999; Crystal, 2006; Virtanen, 2013). Second person present tense can be awkward (Cherny, 1999). This may be so because emotes are self-referential or reflexive (Cherny, 1999; Virtanen, 2013). They are meant to serve as a means of communicating one's behavior including actions, reactions, gestures, and facial expression (Crystal, 2006). Emotes "should be recognized as portraying the speaker in a certain way" (North, 2007, p. 542). North (2007) refers to them as "enactments." Werry (1996) likens them to stage directions that one might find in a script, which are written in third person. For example, in "The Zoo Story" by Edward Albee (1959), some stage directions for the character Jerry include: "*JERRY snorts but does not move*" and "*JERRY laughs, stays*". Emotes are narratives, and the interlocutors is the playwright who writes an autobiographical sketch of him/herself. Thus, emotes are inherently performative (Cherny, 1999; Herring, 2012; North, 2007; Virtanen, 2013; Werry, 1996). They are a

⁵² These are examples of positioning the arrow such that it points to the interlocutor's name, which the chat programs print out to the screen before printing the interlocutor's comments.

“means of foregrounding the structure of the activity-dimension of an interaction” (Crystal, 2006, p. 190). As such, emotes open up conversation to play.

In the conversation below,⁵³ Mike and Jerry make sexual advances toward Lauren, who decides to turn the tables on both of them by playing one off the other. Mike and Jerry respond to her flirtations both conversationally and emotively. When she ** kisses Jerry**, he responds by letting Lauren know that “*that will do.*” Then she rebuffs him and walks back to Mike, playing the part of the sexy devotee. Mike responds by romantically popping grapes in Lauren’s mouth, which sickens Ambie who has just re-entered the room. In response to Ambie’s disgust, Lauren takes her joke further by extending her flirtations to Ambie and offering her a drink.

User1-MIKE: lauren. i just want you in a sexy bikini thats all

User2-JERRY: 8-)

User2-JERRY: i want you in a sexy thong lauren

User3-LAUREN: lmao Jerry

User2-JERRY: hehehe

User3-LAUREN: *** has only sexy bikini on hands mike his beer***

User2-JERRY: and of course i have other ideas

User1-MIKE: thanks babe

User3-LAUREN: your not welcome Mike *** kisses Jerry***

User2-JERRY: o well that will do lauren

User1-MIKE: what about me

User3-LAUREN: your too young Mike

User3-LAUREN: lmao

User2-JERRY: :-X

User3-LAUREN: sorry mike *** walks back and sits near Mike and grabs and budlight bottle with lime in it***

User4-AMBIE re-enters the room after having left briefly.

User2-JERRY: ambie were you in the room my crib yo?

User4-AMBIE: i left [>:o] fkin erin and ray all up on each other

⁵³ Emotes are marked in bold.

User4-AMBIE: [>:o]
User1-MIKE: : **pops some grapes in lauren`s mouth very romantically**"
User2-JERRY: o i was gonna have you tell everyone i said hey please
User4-AMBIE: ok enough please ya'll gonna make me puke
User4-AMBIE: :-\
User2-JERRY: damn :-(
User3-LAUREN: what can i get you to drink Ambie?
User4-AMBIE: the biggest bottle i can break on erin's head
User4-AMBIE: haha
User4-AMBIE: j/k

Crystal (2006) claims that the use or non-use of emotes relies on the character and style of the communication group. Lauren, Mike, Jerry, and Ambie clearly have a shared history. (Only one of them has a userid that contains the name by which they are referred by the others.) It is possible Lauren, Mike, and Jerry have ritualized their flirtation games in the way that many of the participants in Cherny's (1999) MOO routinized their linguistic games. Ambie left the room prior to this conversation and reacts with immediate disgust upon re-entry, based on seeing only one clue that the game is in progress: Mike's emote about feeding grapes to Lauren. This suggests that the game is nothing new. Games such as these require willing collaborators (Cherny, 1999), but Ambie is not willing to play. The game ends when Ambie vocalizes her disinclination to play along, in spite of Lauren's equal opportunity advances.

This conversation demonstrates what Cherny (1999, p. 211) refers to as "emoted byplay," which can produce a "cartoon-like atmosphere." This exchange shares many of the humorous qualities of old Tex Avery cartoons featuring the lecherous wolf, such as "Swing Shift Cinderella" and "Red Hot Riding Hood," or even Robert Zemeckis' movie "Who Framed Roger Rabbit?"

The conversation below provides another example of performativity and collaborative play.⁵⁴ The participants in this conversation fight over a hot dog and a glass of iced tea. It begins with Joe announcing that he is grilling a hotdog and Ms whimpering as he looks on, wishing he had the hotdog. The *whimpers* emote launches, or *keys* (term used by Cherny, 1999), the game in the way that a referee’s whistle starts an athletic match. Joe picks up the ball by telling Ms “*NO*” and “*back off.*” Ms grabs the hotdog, asking “now what?”, as if he didn’t fully think his larcenous maneuver through. Oz jumps into the game by stealing Joe’s hotdog from Ms and throwing it into the pool. Git/Johnnie takes Joe’s side by offering him a bat to keep Oz and Ms off. Joe, giving away too much and thus baiting the line, announces that he also has iced tea. Ms bites—he drinks the tea, eats the hotdog, and produces a victory belch to punctuate his emote. Oz takes the bait as well and stalks Joe across the pool area, grabs the remainder of his iced tea, and dumps it into the pool.

User1-MS: wb joe...when you gonna hold still

User2-JOE: i was grilling me a hotdog

User2-JOE: never ms lol

User3: grill it all and then come back

User1-MS: **whimpers at joes hotdog**

User1-MS: lol

User2-JOE: lol

User2-JOE: NO

User2-JOE: back off ms its mine

User1-MS: **sniffs a bit closer**

User2-JOE: and i intend to enjoy it!

User2-JOE: BACK OFF

User4: hmmm hmmm hot dog n ketchup hmmm yep that sounds good brb

User1-MS: **whimpers...ouch**

User5-OZ: ***grabs joes hot dog***

⁵⁴ This conversation has been edited. All utterances that did not pertain to the hot dog game were removed so that the example would be easier to follow.

User6-GIT/JOHNNIE: ty

User8: iced*

User9: bless u

User5-OZ: ***dumps joes iced tea***

Cherny (1999) says that emoted byplay often includes play with imaginary objects. The hotdogs and iced teas in this conversation are central to the unfolding drama. They change hands in the way that chainsaws, or other props, are passed between a group of street jugglers. Cherny (1999) likens play with imaginary objects to mime, where performers create a rich world out of thin air—or in this case out of keyboard characters.

The hotdog game might appear at first to be an attack on Joe, but it is more likely that the three main actors—Ms, Oz, and Joe—are old friends who have tacitly defined appropriate limits for teasing, and thus know how far they can push. Cherny (1999) explains that some emotes may appear at first glance to be insults, but this is not the case. She cites examples where MOO participants use what she calls “anti-social” emotes in a self-deprecating way to poke fun at themselves after leveling an attack against another participant. Seemingly hostile acts or utterances in the hotdog game are followed by a softening term, to signal that all was meant in jest. For example, when Ms says his knee is cocked and loaded, ready for kicking Joe, he follows it up with *lol (laughing out loud)* to soften the threat. When Joe tells Ms *fu (fuck you)*, he too follows it up with an *lol*. These interlocutors know how to play rough but also know how to help each other up.

In addition to play and performance, emotes may be used for a variety of other purposes including providing feedback (i.e., minimal responses), support, and affection (Cherny, 1999). For example, **nods**, **shrug**, and *giggles* (found in the corpus) enable the interlocutor to signal his/her attention to the conversation. *WINK*, *-hugs tight-*, and *muahs*

(the sound of giving a kiss—an emote that includes onomatopoeic expression) convey affection and offer support.

Cherny (1999) and Herring (2012) claim that some emotes have become conventionalized. Cherny (1999) cites the example of *waves* in the MOO she examined. That particular emote was used infrequently in the corpus. However, variants of *hugs*, *nods*, and forms of kisses (e.g., *muahs*) were used quite a bit.

Repairs, Addressivity, and Compounds / Space Omission

Cherny (1999), Collister (2008), Ferrara et al. (1991), Lindh (2009), and Wutiolarn and Attaprechakul (2012) all discuss the use of punctuation to symbolize the repairs of disfluencies. Cherny (1999) discusses the use of slashes in “find and replace” programming formulas that are used for repairs in the MOO she studied. Some of the interlocutors in Ferrara et al.’s (1991) study used parentheses. In the other studies referenced above, which all analyzed game conversation, asterisks were cited as the punctuation of choice for making repairs. In this corpus, asterisks were also the most frequently used marker of repairs. There was one instance using a dash and two types that did not use punctuation at all. Out of those that used asterisks, most place the asterisk after the correctly spelled word (16 types). Four types placed the asterisk before the repaired word. One type placed asterisks on both sides.

In Werry’s (1996) analysis of Internet Relay Chat, interlocutors sometimes responded to a particular person by prefacing their comment with the person’s name and a colon. Werry (1996) refers to this as addressivity and claims that it is a form of minimal response that signals “active attention and may be used to indicate understanding” (p. 52). He believes they compensate for “the weakened link between sender and receiver” that results from not being

physically proximate to one another (Werry, 1996, p. 52). Few instances of addressivity were found in this corpus, but this may be due in part to the methods (i.e., weeding standard English terms out of context, which would have included any properly capitalized first names). All instances of addressivity that were found followed a formula of an at sign (@) followed by a person's name (e.g., @Laura).⁵⁵ Three were found in the computing email list, two in the transcriptionists' email list. This feature's limitation to these two communities suggests that the use of the @ sign in this manner is a matter of group style.

The vast majority of types that were marked as compounds / space omission were instances of space omission rather than true compounding.⁵⁶ Yongyan (2000) noted similar instances, terms such as *Ihave, iwould, thankyou, alittle, somuch*, etc. Similar terms are found in the corpus, such as *thereviews, themup, somekind, hisnailing, answeyou*, etc. It is likely that these instances result from typing too fast and inadvertently missing the space bar. As such, they could have easily been classified as misspellings/typos.

Additional Examples of Creativity

“Creativity can be defined as the manipulation of language form, in unexpected and yet contextually appropriate ways.” (Tagg, 2009, p. 159). Tagg (2009) speaks specifically about manipulating morphemes to create novel terms. An example of this would be the “attachment of existing affixes to unusual or unorthodox bases” (Rúa, 2007, p.147) as in the

⁵⁵ These instances should not be confused with the use the @ sign in Twitter. The @ sign in Twitter is a way to link to someone else's Twitter page.

⁵⁶ Plag (2003) defines traditional compounding as the combination of two word bases. Eble (1996) explains that they can be written as one word, two words, or separated by a hyphen; patterns usually include a noun plus a noun, an adjective plus a noun, a noun plus a verb, or some word plus a particle (e.g., *printout, download*). These patterns typically do not include pronouns and functional, non-content words. Compounding links two terms that are associated semantically. It does not usually involve blending two neighboring words whose association is more a matter of proximity than semantics.

example *loladins* provided in the Acronyms / Initialisms section of the Discussion. This involves using existing devices in new ways (Rúa, 2007). In this corpus, morphology, typography, and orthography are manipulated in creative ways to form new terms. In addition to terms that have already been discussed, Table 34 lists and explains other terms that demonstrate creativity as well.

Table 34: Examples of creative word-creation.

Word	Definition	Location
<i>/bug</i> <i>/bugged</i>	The first instance is used as a verb, the second an adjective. The interlocutor is asking others to report a problem he found in EverQuest so that game administrators will take his complaint about the problem more seriously. S/he is drawing on the slash command structure found in many games (such as using them for emotes) and turning the word <i>bug</i> into a way to identify something as a problem. This is typographical creativity but also has to do with creating a new meaning and usage for the term.	EQ forums
^	A way to signify agreement with an earlier comment. It co-opts the caret for new purposes and is thus an example of typographical creativity	WoW chat; TS forums
<i>Anti-bullying</i>	A process that opposes bullying. Specifically the interlocutor is recommending an anti-bullying class in school, in the way that some schools had/have classes on avoiding taking drugs. This is an example of morphological creativity.	Teenspot forums
<i>beware!</i>	This means beware but an extra e is added to pull in the word <i>bee</i> , for humorous effect. This is an example of orthographic creativity.	Beekeepers' email list
<i>boomkin</i>	In WoW, one can play a druid and specialize his abilities toward a "Balance" profile. When using these abilities, druids shapeshift into "Moonkin form" which makes the druid look like a giant owl. Moonkins can perform nature spells like calling down thunderstorms. These spells go "boom" and so <i>Moonkin</i> was transformed into <i>boomkin</i> . This is both a blend and figurative. It qualifies as morphological creativity.	WoW chat
<i>butthurtness,</i>	The state of one's butt hurting. It uses affixation and is an example of morphological creativity.	Teenspot forums
<i>buzzification</i>	The act of turning something into a buzz word. This is an example of morphological creativity.	Computing email list
<i>catwalking</i>	Walking like a cat. This is a compound and an example of morphological creativity.	SMS
<i>cloud-in-a-box</i>	This is a compound used to refer to a product that would enable someone to set up a cloud-computing platform. Thus it is an example of morphological creativity.	Computing email list
<i>couchsurf</i> <i>couchsurfing</i>	To watch television while on the couch. This is a compound and is thus an example of morphological creativity.	Transcriptionists' email list

Word	Definition	Location
<i>de-D&D-ization</i>	The process of making something not like Dungeons and Dragons. This is creative affixation, and thus an example of morphological creativity.	Gamers email list
<i>de-Tolkienization</i>	Eliminating Tolkien influences or references from a work. This is an example of affixation, and thus morphological creativity.	Gamers email list
<i>don';-)t</i>	An emoticon inserted into <i>don't</i> to soften the negativity of the statement. This is an example of infixation, and as such is an example of morphological creativity.	SMS
<i>e-Face</i>	A person's online presence. This is affixation and thus morphological creativity	Teenspot forums
<i>e-people</i>	People you may only know online. This is affixation and thus morphological creativity.	Teenspot forums
<i>Eggseolent</i>	Excellent. This is an example of orthographic creativity.	SMS
<i>faceroll</i>	A derogatory term for an action that was poorly executed, almost as if the gamer had rolled his face across the keys instead of carefully selecting the right keys to call up the right character abilities to suit the situation. It is a compound, and as such is an example of morphological creativity.	WoW chat
<i>goshness</i>	The state of being astonished. It is an example of affixation, and thus morphological creativity.	UNC IM
<i>grab-and-go</i>	Types of quests that do not involve a lot of thinking or skill. It is a compound, and thus an example of morphological creativity.	WoW forums
<i>happ-bee</i>	<i>Bee</i> is added to <i>happy</i> to make a play on the word <i>bee</i> . This is a blend, and thus morphological creativity.	Beekeepers' email list
<i>lawl</i>	<i>Lol</i> phonetically spelled. This is an example of orthographic creativity.	Teenspot forums
<i>loladins</i>	Those who laugh out loud. This is affixation to an acronym, and as such is an example of morphological creativity.	WoW chat
<i>Lulz.</i>	<i>Lols</i> (lots of laughing out loud) phonetically spelled. This is an example of orthographic creativity.	EQ forums
<i>mobo</i>	This is a rhyming shortening of <i>motherboard</i> . It could be considered orthographic creativity.	Yahoo forums
<i>ragequit</i>	To quit playing the game because one feels so much rage toward it. This is a compound, and as such is an example of morphological creativity.	WoW chat
scrollin'scrollin'.....scrollin'....keep them words a rollin'.....CH ATHIDE!!!!!!!!!!!!!!!!!!!! !!	This is a spoof of the "Rawhide" song for chat. The ellipses are added to simulate the rhythm of the song. This is an example of morphological and typographic creativity.	NPS chat
<i>sesky</i>	<i>Sexy</i> respelled for the purpose of playing with the sounds. This is an example of orthographic creativity	NPS chat
<i>sweeth;-)eart.</i>	An emoticon inserted into <i>sweetheart</i> . This is an example of infixation, and as such, is an example of morphological creativity.	SMS
<i>Threadstarter,</i>	Refers to the person who started the forum thread. This is a compound and as such, is an example of morphological creativity.	Teenspot forums

Conclusion

Crystal (2006, 2008b) discusses popular press accounts of the fear some have about cyberlanguage signaling the demise of “proper” English. He attempts to allay these fears by explaining that many of the features used in the creation of new terms online are “centuries old” (Crystal, 2008b, p. 27). The percentage of cyberlanguage tokens out of all tokens collected for this corpus provide evidence to support Crystal’s attempts at amelioration. Although new terms are being created frequently, they do not appear in overwhelming proportions. The 14,681 cyberlanguage tokens found in this corpus comprise a mere 10.75% of all tokens in the corpus.

Features are not evenly distributed, but if they were, the rate at which they appear is small. The three most-frequent features in the corpus were acronyms/initialisms, non-standard use of lowercase, and shortenings (see Table 19). A little over 20 acronyms appear per 1,000 words; 19.09 lowercase tokens appear per 1,000 words; and 11.84 shortenings appear per 1,000 words. Some of the least frequent features in the corpus were repairs, pointing, addressivity, and pictograms. All of these appear less than 1 time per 1,000 words (0.18 repairs, 0.14 pointing instances, 0.04 forms of addressivity, and 0.01 pictograms).

Shortis (2007, p. 17) claims that cyberlanguage features are not used with reckless abandon; most of the time online conversation follows a “standard English default.” Use of any variety “is a matter of appropriateness and identity rather than a matter of rectitude and uniformity” (Shortis, 2007, p. 17). People choose their words to match the context. Online, it is appropriate to use abbreviations and surrogate face-to-face cues. Not to do so limits communication. In some situations, speed and brevity is of the essence, and thus one’s ability to keep up with the conversation is key. In most online situations, interlocutors cannot see

and hear one another. When ordinary words are insufficient for conveying sentiment, rapport, social meaning, camaraderie, and jest, interlocutors must avail themselves of all the tools available to them and they must exploit them to their highest potential to successfully convey their thoughts and emotions. These tools include the range of characters on a computer keyboard, and exploiting them means that some conventions, some rules of “proper” spelling and grammar, must be thrown out or bent in some way. Reshaping English to suit the communication situation does not necessarily mean that interlocutors have poor understanding of English. In fact, Crystal (2008b) would argue otherwise. For example, vowels, more so than consonants, are removed from words to shorten them. Crystal (2008b) explains that this is because interlocutors realize that consonants carry more weight, and omitting them may exact costs to intelligibility. Many of those who use cyberlanguage features to bridge the gap between themselves and their fellow conversants do so with linguistic mastery. Writing is best understood in terms of social functions and context (Shortis, 2007). The revisions to English discussed here achieve social and utilitarian goals, and not enough of them exist or are adopted into general language to cause alarm. English is just fine online and off.

Conclusion

This research study was conducted over four years. In the course of this work, the researcher produced a corpus of 136,529 tokens (23,912 types) that spans five types of online, conversational media (forums, email, SMS, IM, and chat), from 17 different sources (Teenspot.com forums, Yahoo! forums, EverQuest forums, WoW forums and chat, Dr. Sotillo's SMS corpus, UNC Ask a Librarian IM, NCKnows IM, L-net IM, AOL chat, NPS chat, and six email lists), covering a variety of topics (Gaming, Technology, and Other) and purposes (Serious and Recreational/leisure-oriented). It includes conversations that occurred over a nine-year period (2003-2011). All 136,529 terms were manually analyzed for the presence of 28 cyberlanguage features identified by other researchers as well as three additional features uncovered in this analysis. Most of these tokens—121,848—were weeded because they were general/standard English. The cyberlanguage tokens that remained totaled 14,681 (6,604 types). To the researcher's knowledge this is the first corpus of this kind, covering this much breadth of online communication.

The goal of this research was to test assertions made by other researchers about the influences cybermedia and other situational variables may have on language production, and thereby assess the validity of the notion that technology exerts influences on communication behavior. Specifically, it aimed:

- to compare language—specifically cyberlanguage feature frequencies—across media, specifically, forums, email, SMS, IM, and chat,

- to compare language—specifically cyberlanguage feature frequencies—across situational factors (or genre factors), such as topic and purpose, and
- to identify examples of linguistic creativity

so as to provide a more comprehensive description of cyberlanguage as it appears across online media. Three research questions guided this analysis:

RQ 1: What cyberlanguage features are common across online, conversational media and genre situations?

RQ 2: What cyberlanguage features differ between media and genre situations and how do they differ?

RQ 3: Are there examples of linguistic creativity that may serve as evidence of an interlocutor's ability to respond and adapt to technological change in innovative ways? If so, what are some examples?

These goals were accomplished and the questions were answered.

Terms in the 139,529-word corpus were evaluated for the presence of cyberlanguage features. If features were present, the term was classified according to the feature or features it contained. Statistical tests, namely chi-square tests, were used to compare frequencies, and results were interpreted in light of specific examples in the corpus and discoveries made by other researchers.

Table 32 along with the insignificant chi-square values listed in Tables 21 through 31 answer research question 1: *What cyberlanguage features are common across online, conversational media and genre situations?* Few features were distributed in a homogeneous fashion, and they were homogeneous in only topic and purposes comparisons, not in media characteristic comparisons. These features include number homophones, onomatopoeic expression, letter duplication, misspellings and typos, affixation and use of combining forms, and compounds/space omission.

More variety in feature use and term creation, however, was unearthed in these results. Significant chi-square values were returned for all features for which chi-square tests were appropriate when features were compared across the five media. This shows that feature use varies based on medium, and that medium is a good discriminator of online language variation. Table 33 summarizes the differences between features, and along with Tables 20 through 31, answers research question 2: *What cyberlanguage features differ between media and genre situations and how do they differ?*

There is validity to the notion that technology exerts some influence over communication behavior, because of the differences found in feature frequency in their comparison across media factors. This is not to suggest that interlocutors are at the mercy of technology and act in completely deterministic ways. Rather, interlocutors recognize a medium's advantages and disadvantages, and because they desire conviviality and social connection, they creatively and masterfully find ways to bend the typography, orthography, and morphology of written language to achieve these ends in spite of any medium-based deficiencies. Additionally, other contextual variables or genre factors, as they are called in this paper, were shown to be associated with differences in feature use. Both medium and genre conjointly form contexts in which interlocutors use abbreviations, surrogate face-to-face cues, and other features atypical of traditional writing.

In addition to media and genre influences, some of the variation found in feature use can be attributed to the individual style of particular interlocutor groups. Participants in the opera fan email list substitute asterisks for quotation marks; and AOL and NPS participants decorate usernames when greeting their friends. Terminology is also developed in some

communities to help them describe their activities. For example, WoW players have created an acronym-heavy vocabulary around the trading and selling of game items.

Many terms and phrases are created to establish a modality of play and performance. Surrogate face-to-face cues, such as emotes, phonetic respellings, and onomatopoeic expressions, are particularly well-suited to play. Emotes, especially, are able to bring interlocutors together in an impromptu performance of self through text. Emote-saturated games like those described in the Emotes chapter, acronyms that wish others well, simulated accents, and plays on sound are just some of the many ways that interlocutors communicate phatically with one another, offering support and care, lightening the mood, and signaling attention to the conversation.

Unlike those terms and phrases used for play and performance, many terms and features are quite ordinary. They are used in high frequency and seem to be online-communication staples. *Lol*, onomatopoeic laughter (e.g., *haha*), the smiley emoticon (:) , and variants on the *hugs* emote are a few examples of common terms. Because these have outworn their novelty, some, such as North (2007), would not consider them creative. Similarly, acronyms and lowercase are common features that may also be online-communication staples.

Creativity is demonstrated by terms whose original form has been manipulated in unexpected ways (Tagg, 2009). *Loladins* (those WoW players who laugh out loud) and *bewbs* (playfully emphasizing the *oo* sound in *boobs*) are two examples of creativity. “Creative co-construction of relationships” through language play and games are yet another way interlocutors can creatively shape their communication online (Carter, 2004, p. 188). The collaborative games described in the Emotes chapter are an example of this kind of

creativity. Artistic renderings of typography are another creative act. Some emoticons, such as :-*:-* (a kiss reduplicated for emphasis, cartoony in its use of asterisks to symbolize puckered lips), and username decoration, such as »©«*''`*»@}~©«*Hey Hey Howdy Angie*»©~{@«*''`*»©« are examples of this kind of pictographic creativity. These and other examples discussed in the Discussion section answer research question 3: *Are there examples of linguistic creativity that may serve as evidence of an interlocutor's ability to respond and adapt to technological change in innovative ways? If so, what are some examples?*

Most of the cyberlanguage terms in the corpus were hapax legomena (5,211 or 78.91% of cyberlanguage types), and an even greater number contained fewer than 10 tokens (6,450 or 98%). Only two percent occurred 11 or more times. The corpus includes texts written over a period of nine years (2003-2011). That is an average of 579 hapax legomena per year for only a small set of Internet conversations. This suggests that language is changing rapidly online, and several innovations that are not possible in speech (e.g., emoticons, pictographic offsetting, plays on orthography) and not likely in standard forms of writing were identified. But whether this change is more or less than change in other language forms cannot be inferred from these results. This is especially so in the case of everyday speech which is most often not captured in the way that online communication is, so there is little to no data with which to make such comparisons.

Out of the 18,245 cyberlanguage feature identified, 56.38% (10,287) were some sort of abbreviation; 31.38% were some sort of surrogate face-to-face cue, and 12.24% were other types of features (those listed toward the end of Table 2).

Table 35: Frequency and proportion of types of features.

Feature Type	Frequency	% of All Features (N = 18,245)	% of All Tokens (N = 136,529)
Abbreviations	10,287	56.38%	7.53%
Surrogate face-to-face cues	5,725	31.38%	4.19%
Other features	2,233	12.24%	1.64%

This suggests that brevity and speed are important when conversing online. The 31.38% of surrogates, however, should be considered along with the many general/standard English terms that were weeded initially. For example, many emoted phrases include general/standard English terms, and the kind of games and language play shown in the Emotes chapter would not be possible without the use of general/standard English terms. Thus, creating a sense of social presence involves using both surrogates and general terms conjointly. Some abbreviations, such as acronyms symbolizing laughter, also help to fulfill a surrogate face-to-face cue role. Thus, the number of surrogates is probably higher. So although the percentage of surrogates appears to be lower than that of abbreviations, they still play an important role in enabling users to close the distance between them.

When considering the number of cyberlanguage features in light of the total number of terms collected, it is clear that the majority of online language is general/standard English. So any fears about the demise of English are unfounded. Interlocutors use cyberlanguage features in creative, playful, and innovative ways, but they do not do so, or do not intend to do so, to the exclusion of clarity.

Future study of cyberlanguage should include the acquisition and analysis of additional text samples from different media and communities, to continue to expand the description of cyberlanguage presented here. The corpus currently covers conversations from

2003-2011, and thus potentially affords investigations into language change over time, including hapax and other term adoption, trends in term usage, term decay, and new and emerging features. The addition of newer conversations will only make such investigations richer. The range of sources should also be expanded, perhaps to include sources that—for this study—did not fit the definition of conversation as closely as was desired, such as Twitter and Facebook wall postings. These sources are becoming increasingly popular and would afford comparison of more monologic, broadcast communication with true conversation. Further additions could include conversations from other kinds of topics, purposes, and user groups, such as attempting to find other kinds of IM sources besides virtual reference and additional SMS messages.

The corpus in its current state provides a wealth of opportunities for deeper investigations focused on specific topics; additions to the corpus would only increase these opportunities and result in richer findings. These investigations could focus on the features that were not discussed in detail in the Discussion section, or on delving deeper into novel usages noted in this dissertation, such as non-emphasis-marking offsetting.

This study focused primarily on lexical analysis, but the corpus created for this analysis could also be used to analyze communication at the level of discourse. Other investigations could center on use of cyberlanguage for different communication purposes, such as (a) establishing rapport and creating intimacy or, conversely, distancing one's self from others through hostile language usage (using communication accommodation theory⁵⁷ as a backdrop), (b) establishing covert or over prestige through the use of cyberlanguage

⁵⁷ Cf. Giles, H., Coupland, J., & Coupland, N. (1991). Contexts of accommodation: developments in applied sociolinguistics In *Accommodation theory: Communication, context, and consequence* (pp. 1-68). Cambridge: Cambridge University Press.

(using William Labov's work⁵⁸ on the subject as a backdrop), (c) playing with identity and language through the use of ritualized language games (using Erving Goffman's work⁵⁹ on self-presentation and forms of talk, and Crystal's and Danet's work⁶⁰ on language play), as well as other purposes. Other investigations could focus on more in-depth, smaller-scale comparisons of the themes found in this study (i.e., play, performance, and creativity). For example, a comparison of play, performance, and creativity in different kinds of chat sources, such as comparing more general chat, as in AOL chat, with MMOG chat, as in WoW, would help answer questions about the tenor of gamer conversation. Are gamers as serious as Yee's (2006) work suggests?

Finally, a lexicon of cyberlanguage terms and list of rules for automatically detecting any that can be automatically detected would be a significant contribution to the information science community, in particular to those working in the areas of information retrieval and natural language processing. The lexicon could continue to grow as the corpus grows. This would enable information scientists to extend their efforts to communication and information arenas that may be receiving little attention at this time. Cybermedia are used not only for recreation and frivolity, but also serious purposes. As such, they may contain valuable information that would be more successfully mined with the aid of a lexicon of cyberlanguage and rules for its automatic detection.

⁵⁸ Labov, W. (2006). *The social stratification of English in New York City* (2nd ed.). Cambridge; New York: Cambridge University Press.

⁵⁹ Goffman, E. (1981). *Forms of talk*. Philadelphia: University of Pennsylvania Press; and Goffman, E. (1990). *The presentation of self in everyday life*. N.Y.: Doubleday.

⁶⁰ Such as Crystal, D. (1998). *Language play*. Chicago: The University of Chicago Press; and Danet, B. (2001). *Cyberpl@y: Communicating online*. Oxford: Berg Publishers.

Appendices

Appendix A: A Note about the Term *Cyberlanguage*

There are several popular terms that describe online communication, such as computer-mediated communication, computer-mediated discourse, and electronic communication (used by Herring, 2002, 2001, and 2012 respectively). Some researchers have coined terms such as Netspeak (Crystal, 2006) and Interactive Written Discourse (Ferrara, Brunner, and Whittemore, 1991), and use of specific labels is criticized (Herring, personal communication, 2012). For example, Dürscheid (2004) and Dürscheid and Frehner (2013) criticize Crystal's use of Netspeak because they believe it to be a generalized summation of what is in reality a variety of text types. They explain that the term implies a “new, previously unknown language with unique features, thus deserving its own term” (p. 41). However, Crystal (2006, p. 271) questions Netspeak as a homogeneous language, and instead suggests it is a collection of distinct dialects. In his book on SMS, Crystal (2008b, p. 27) points out that the linguistic processes that are used are “centuries old,” and he later explains that genuine novelty exists in the ways that interlocutors take these age-old processes and stretch them beyond their original, more conventional, uses.

Although Dürscheid and Frehner's (2013) criticism of using a specific “terminus technicus” to describe online communication is accepted in the research community, there are several reasons why online language—or Netspeak or cyberlanguage—is worthy of its own special label.

(1) Although some cyberlanguage features discussed in the Literature Review may have originated outside of the Internet, they are used more frequently online than in

general/standard English texts, as documented in several of the papers cited in this dissertation. Shortis (2007, p. 18) explains that online/digital communication venues have “revolutionized what counts as spelling by legitimizing and popularizing longstanding vernacular orthographic practices found in popular and domestic culture but underrepresented in public, academic and media accounts of language use, and in linguistic corpora, which largely draws from texts spelt in standard English.”

(2) Some of these features (or word-creation processes, if you will) have been reengineered to suit the online situation, making their reinvention closely tied to the special nature of online communication (e.g., using asterisks for repairs, using punctuation and sometimes letters to convey facial expression as in the case of emoticons).

Among the most obvious of developments occasioned by the explosion of information and communication technologies in the past twenty years is the rapid increase in the lexicon, as new words appear which refer to new communicative activities. Among the most striking innovations are those in which a basic form is creatively extended into a range of new formations and contexts. (Carter, 2004, p. 189)

(3) New terms are introduced frequently through the use of these features, as shown by the large number of low-frequency terms and hapax legemenona in the corpus used for this dissertation study (5,211 hapaxes or 78.91% of the 6,604 cyberlanguage word types). (4) Some features were born specifically out of online conversation, such as the emoticon, which is credited to a forum posting made by Scott Fahlman, a computer scientist at Carnegie Mellon (Dresner & Herring, 2010).

(5) Furthermore, Crystal’s critics—namely Dürscheid—have coined their own labels (e.g., *keyboard-to-screen communication* in Jucker & Dürscheid, 2012), and who is to say

which label is best? Jucker and Dürscheid (p. 40) explain that their term is less problematic than those that use the word *computer* (such as *computer-mediated communication*) because it omits *computer*, which they believe is important because “cell phones are usually not considered to be computers.” This researcher—coming from an information technology background—would argue that cell phones are most definitely computers. They have operating systems and programmed applications that allow users to execute commands to perform tasks. Just because a cell phone is very small makes it no less a computer. Furthermore, some modern cell phones do not possess keyboards as input devices, and instead offer users touch-screen interaction, making the term *keyboard* as problematic as Jucker and Dürscheid (2012) claim *computer* is. The choice of label may simply be a matter of personal preference or viewpoint, making criticisms of specific labels an unproductive exercise.

(5) And finally, having a label of some kind is necessary, because not having a label makes reference to the phenomenon in a written work or conversation awkward and confusing. So for the purposes of this paper, online conversational language will be referred to as *cyberlanguage*. The prefix *cyber-* was selected because it refers to early Internet days and hacker culture. Crystal (2006, pp. 74-75) explains that the future of this language is “very much bound up with the extent to which hacker-originated language and style has developed a sufficiently stable and powerful identity to motivate new Internet users to use it.” Hackers are digital cowboys reminiscent of William Gibson’s character Case in the novel “*Neuromancer*,” and William Gibson coined the term *cyberspace*, from which the prefix *cyber-* is borrowed. According to the Jargon File (<http://www.catb.org/jargon/html/index.html>), hackers are creative people who enjoy

wordplay; thus this label was selected because it honors online language's early roots and its playful character.

Appendix B: Differences between Speech and Writing

Speech	Writing
More dialogic. (Baron, 2010; Carter, 2004; Hård af Segerstad, 2002)	More monologic. (Baron, 2010; Carter, 2004; Hård af Segerstad, 2002)
Ephemeral. As time proceeds, interlocutors “can no longer observe that which was produced earlier.” (Baron, 2010; Hård af Segerstad, 2002, p. 43)	Durable. It is “persistent and may be reread and stored for the future.” (Baron, 2010; Hård af Segerstad, 2002, p. 43)
Multiple channels are used (e.g., auditory, visual) so face-to-face signals are possible. (Baron, 2010; Hård af Segerstad, 2002)	Monomodal, so face-to-face signals are not present. (Baron, 2010; Hård af Segerstad, 2002)
Interlocutor physical characteristics (e.g., ethnicity, sex, age) are more obvious. (Hård af Segerstad, 2002)	Interlocutors physical characteristics are obscured.
Interlocutors are present at the same time and place so they share context. (Carter, 2004; Hård af Segerstad, 2002)	Context may not be shared by interlocutors. Writing the text is independent of the situation in which it is read. (Hård af Segerstad, 2002)
Because of shared context, interlocutors may be more implicit and do not have provide as many explanatory details to mark the context. (Hård af Segerstad, 2002)	“The absence of immediate context must be compensated for, i.e. referents must be fully described, and arguments must be represented more extensively.” (Hård af Segerstad, 2002, p. 43)
“Shorter units of expression” (Baron, 2010, p. 2)	“Longer units of expression” (Baron, 2010, p. 2)
Less structural complexity, less lexical density, and less varied vocabulary. (Baron, 2010; Hård af Segerstad, 2002)	More structural complexity, higher lexical density, and more varied vocabulary. (Baron, 2010; Hård af Segerstad, 2002)
Vocabulary is more likely to be concrete, colloquial, and to exhibit fewer abbreviations, but more slang and obscenity. Pronouns tend to be first and second person. (Baron, 2010)	Vocabulary is more likely to be abstract, literary, and to exhibit more abbreviations, and less slang and obscenity. Fewer first and second person pronouns tend to be used (except in letters). (Baron, 2010)
By and large, unplanned, unrehearsed, spontaneous. Has more disfluencies, pauses, false-starts, self-corrections. (Carter, 2004; Hård af Segerstad, 2002)	Time to plan and edit so the final version appears polished. (Hård af Segerstad, 2002)
Creative expression is co-created, organic, and contextual. It involves interpersonal interaction and contains more overt markings of attitude. It is more representational, expressive, non-literal, affective, and playful. (Carter, 2004)	Creative expression is more rule-governed, referential, literal, and serious. (Carter, 2004)

Appendix C: Support for Topic and Purpose Classifications

To disguise sources—and thus participants—as much as possible, the evidence below is discussed generally.

Source (and specific forum for forum sources)	Topic	Purpose
Teenspot: General	<p><i>Other</i></p> <p>The name of the forum happened to match the Other topic classification. The tagline for the forum encouraged talking about anything and everything. Because there were forums dedicated to gaming and technology, it was assumed gaming and technology topics would largely be discussed in those forums, and that Other topics would have been discussed in the General forum.</p>	<p><i>Mixed</i></p> <p>The tagline suggested the possibility of a variety of purposes.</p>
Teenspot: 1/2 School and 1/2 College	<p><i>Other</i></p> <p>The names of the forums suggested that they were intended for discussion of school and college.</p>	<p><i>Mixed</i></p> <p>The tagline for the School forum encouraged talking about teachers or getting help with homework. The College forum was vague, encouraging Teenspotters to discuss college if they have something to say about it.</p> <p>Talking about one’s teachers seemed a bit like gossip and thus suggested non-serious purposes, but getting help with homework fell into the serious purpose classification. The college forum tagline was non-specific. The most likely purpose based on these would be mixed—i.e., both serious and recreational may be possible.</p>

Source (and specific forum for forum sources)	Topic	Purpose
Teenspot: Technology	Technology The name of the forum suggested that it was intended for discussion of technology.	Ambiguous The tagline describing this forum described technology as being all around us and controlling the world. This provided no clues as to purpose, and participants may have chosen to use the forum for serious purposes (e.g., help with a software application needed for homework) or recreational purposes (e.g., comparison of music playlist sharing programs for personal enjoyment).
Teenspot: Gaming	Gaming The name of the forum suggested that it was intended for discussion of gaming. The tagline describing this forum encouraged participants to post questions and tips on defeating game bosses advancing characters through the game, which confirmed the gaming focus.	Recreational Gaming is recreational.
Yahoo: Schools & Education	Other Yahoo described this forum as a place to discuss colleges, universities, homework, and other school-related issues.	Serious With all Yahoo forums, there was no discussion of the purpose of the forums. However, given the topic description, it could be inferred that the purpose fits this study's definition of serious.
Yahoo: Family & Home	Other Yahoo described this forum as a place to discuss families, genealogy, homes, gardens, etc.	Recreational With all Yahoo forums, there was no discussion of the purpose of the forums. However, given the topic description, it could be inferred that the most likely purpose fits this study's definition of recreational/leisure-oriented in that it is not work or school-related.

Source (and specific forum for forum sources)	Topic	Purpose
Yahoo: Computers & Internet	<p>Technology</p> <p>Yahoo described this forum as a place to discuss cyberculture, hardware, the Internet, etc.</p>	<p>Ambiguous</p> <p>With all Yahoo forums, there was no discussion of the purpose of the forums. Even with the topic description, it is difficult to infer a purpose. It is possible this forum could have been used for serious purposes (e.g., fixing a computer used for work) or recreational/leisure-oriented purposes (e.g., a computer-building hobby).</p>
Yahoo: Games	<p>Gaming</p> <p>Yahoo described this forum as a place to discuss board games, card games, and computer and video games. A subforum about computer and video games was selected.</p>	<p>Recreational</p> <p>With all Yahoo forums, there was no discussion of the purpose of the forums. However, given the topic description, it could be inferred that the purpose is recreational.</p>
EverQuest: Gameplay Mechanics	<p>Gaming Technology</p> <p>At the time of data collection, this forum was described as being a place to discuss user interface issues and other facets of how the game functions. It was used to discuss the game software itself and how to get help using it. For example, some of the threads included discussions about the game crashing, game settings, user interface, game updates, etc.</p>	<p>Recreational</p> <p>At the time of data collection, this page: http://corporate.station.sony.com/en/about-soe.vm described Sony Online Entertainment, the maker of EverQuest, as a designer of "games" that are "designed to push the envelope of online entertainment quality." This suggested that all forums would be intended, ultimately, for recreational purposes.</p>
EverQuest: Gameplay Content	<p>Gaming</p> <p>At the time of data collection, this forum was described as being a place to discuss quests, raids, titles, game art, etc. It was intended for discussion about the game world and how to play the game.</p>	

Source (and specific forum for forum sources)	Topic	Purpose
EverQuest: The Newbie Zone	<p>Gaming</p> <p>At the time of data collection, this forum was described as being a place for new players and players who had left the game and recently returned. It appeared to be the place for new or returning players to get help and become acquainted or re-acquainted with how the game was played. For example, some threads included discussion on choosing and creating a character.</p>	
EverQuest: The Veteran's Lounge	<p>Gaming</p> <p>At the time of data collection, this forum was described as a place for long-time, experienced players to discuss game play. So more advanced game play topics were discussed than in The Newbie Zone.</p>	
World of Warcraft forums: 1/2 Technical Support and 1/2 Mac Technical Support	<p>Gaming Technology</p> <p>At the time of data collection, the taglines describing these forums encouraged discussion about problems with game installation, patching, connecting to servers, crashing during game play, etc. This suggested discussion of the game software itself and how to get help with problems using it.</p>	<p>Recreational</p> <p>At the time of data collection, this page: http://us.blizzard.com/en-us/company/about/ described Blizzard Entertainment®, the maker of World of Warcraft, as “a premier developer and publisher of entertainment software.” The page went on to explain that “by focusing on creating well-designed, highly enjoyable entertainment experiences, Blizzard Entertainment has maintained an unparalleled reputation for quality since its inception.”</p>
World of Warcraft forums: General	<p>Gaming</p> <p>At the time of data collection, the tagline for this forum encouraged discussion of various World of Warcraft game world and game play topics. Threads included discussions of character class abilities, mounts (things characters can ride—like horses), and quests. All these are related to game play.</p>	<p>On http://us.blizzard.com/en-us/company/about/mission.html, Blizzard was further described as “dedicated to creating the most epic entertainment experiences...ever.”</p> <p>This suggested that the intended purpose of World of Warcraft was to provide entertainment, thus recreation.</p>

Source (and specific forum for forum sources)	Topic	Purpose
World of Warcraft forums: New Player Help and Guides	Gaming At the time of data collection, the tagline for this forum described the forum as being a place for new players to discuss the game and get help from more experienced players with game play. Threads included discussion about setting up mentoring programs for new players, questions about which gear/items to purchase, help for character class play, questions about monsters to fight, etc. This forum was akin to The Newbie Zone in EverQuest and was focused on helping newer players get with game play.	
World of Warcraft forums: Quests	Gaming At the time of data collection, the tagline for this forum encouraged players to discuss quests. Completing quests is a core game play activity.	
A multiple sclerosis support group	Other The group was described as being for those who live with multiple sclerosis. Thus it was assumed that the primary topic of conversation was multiple sclerosis, an Other topic.	Serious No information about purpose was provided. However it was assumed to be serious because living with a potentially debilitating medical condition is no small matter. Although some discussions may introduce levity, dealing with such a condition is not superficial, light, or related to amusement. It may be solemn.
A group of transcriptionists	Other The group was described as a communication venue for members of a transcription and proofreading team. Thus the topic was about transcription and proofreading, an Other topic.	Serious Investigations into the larger organization—of which the transcription group was a part—revealed a project-based working structure. Specific work functions were outlined in a detailed flow chart. This suggested a work-oriented focus.

Source (and specific forum for forum sources)	Topic	Purpose
Fans of an opera singer	<p><i>Other</i></p> <p>The title of the email list suggested it was intended for discussion by fans about an opera singer, an Other topic.</p>	<p><i>Recreational</i></p> <p>The group was described as being dedicated to a well-loved, accomplished opera singer, well-known throughout music history. Although nothing explicit was stated about purpose, it was clear this was a fan list where members shared discographies, recordings, videos of performances, photos, and more. This made it “hobby-like” and thus was classified as recreational.</p>
Beekeepers	<p><i>Other</i></p> <p>The group is described as being a communication venue for discussing beekeeping and bees in a west-coast county, thus the Other topic.</p>	<p><i>Ambiguous</i></p> <p>Because it was possible that the participants could be using the list for hobby beekeeping and small business beekeeping, yet no information was provided to confirm one or the other of these purposes, it was classified as ambiguous.</p>
Computing experts	<p><i>Technology</i></p> <p>The name of the email list suggested it was intended for discussion of technology topics related to a particular aspect of computing (which is not disclosed here to protect the privacy of list members).</p>	<p><i>Serious</i></p> <p>The group is described as planning annual conferences about this particular aspect of computing, including issues related to incorporating this type of computing into business models, providing industry experience with this particular aspect of computing, legal issues, and research and development of solutions related to this particular aspect of computing. This suggested serious purposes—legal, business, research, etc.</p>
Multiplayer game players	<p><i>Gaming</i></p> <p>The group is described as a chapter of players of a particular multiplayer game. It is a fantasy role-playing game that allowed players to host multiplayer game instances on personally-owned servers.</p>	<p><i>Recreational</i></p> <p>No information was provided about the purpose of the group, but given that discussion was intended to focus on a game, it was assumed that the purpose was recreational.</p>

Source (and specific forum for forum sources)	Topic	Purpose
Dr. Susana Sotillo, Montclair University	<p><i>Other, Mixed</i></p> <p>Email exchanges with Dr. Sotillo revealed that the messages in her SMS corpus were contributed by 59 individuals—teenagers up to older adults. Messages focused on school-related, personal, and work-related purposes and topics.</p>	
University of North Carolina at Chapel Hill: Ask a Librarian	<p><i>Other</i></p> <p>No specific information was given on the service’s website about topic. However it was inferred, based on setting (i.e., a university), that most topics would be related to school/research-focused information seeking, thus the Other topic.</p>	<p><i>Serious</i></p> <p>No specific information is given on the service’s website about purpose. However, the software used for the service—LibraryH3lp—was developed by the librarian, Pam Sessoms (and her husband Eric), who gave the researcher the IM conversation files. So the website for the software was reviewed. At the time of data collection this website explained that the software was developed to support library virtual reference services at Duke University, North Carolina State University, and the University of North Carolina at Chapel Hill. This goal of supporting library services at academic institutions suggested serious purposes.</p>

Source (and specific forum for forum sources)	Topic	Purpose
NCKnows, North Carolina	<p><i>Other</i></p> <p>No specific information was given about topic on the NCKnows website. However, on http://ncknows.org/using.htm, the following descriptive help text was provided, which alludes to a variety of Other topics: "What sort of information can you provide? We can provide:</p> <ul style="list-style-type: none"> ▪ Facts and Statistics "What is the average income of North Carolina?" ▪ Contexts and Background. "I need to find information on Charles Dickens." ▪ Research Strategies. "I'm doing a paper on affordable housing, where do I start?" ▪ Resources. "I need scholarly articles on schizophrenia and the Internet isn't very helpful, can you help?" ▪ Other information like book recommendations, referrals to experts, information about library holdings and more." 	<p><i>Mixed</i></p> <p>On http://ncknows.org/aboutnc.html, NCKnows is described as "a service that allows North Carolina residents to get help from librarians and use their library resources remotely through a computer. By 'chatting' online with a librarian, you can get the most from your library, including access to articles, audiobooks and more from NC LIVE. Whatever you need, NCKnows will be able to get you started. It's free, helpful and easy. We've helped thousands of NC patrons over the years, including k12 students, business information seekers, college students, people looking for good books and many many more. NCKnows is staffed by librarians from academic, public and specialty libraries. By coordinating with participating libraries across the state, we are able to offer reference help 24/7 except for Sat/Sun midnight to 8 AM."</p> <p>This suggested that the service could be used for serious purposes—e.g., "business information" and "college students"—as well as recreational/leisure-oriented purposes—e.g., "looking for good books."</p>

Source (and specific forum for forum sources)	Topic	Purpose
<p>L-Net, Oregon</p>	<p><i>Other</i></p> <p>No specific information is given about topic. However, on http://www.oregonlibraries.net/services_schools#services, the service is described as being able to “help students with research on the Internet, searching the library catalog, and using online databases such as EBSCOHost. For example: Who was Hernando de Soto? What did Europeans eat in the middle ages? Do video games make people violent? Why is it that lizards are smaller than dinosaurs were? L-net librarians can also help students find short answers to factual questions. I need a map of the country Georgia. What is the population of Burns, Oregon? How long is the Columbia River?"</p> <p>Although one of the questions listed is about gaming, it is not about game play per se. Rather it is a more serious, research question. Thus based on the quote above, this source was classified as being about Other topics.</p>	<p><i>Mixed</i></p> <p>On http://www.oregonlibraries.net/staff/docs/service_guidelines.shtml, L-net librarians are described as being required to chat with patrons “in a friendly and professional manner designed to make the patron feel at ease.” The use of the word professional may suggest serious purposes.</p> <p>On http://www.oregonlibraries.net/for_libraries.shtml, purpose was suggested by this quote: "Get help with research, finding a book or article or verifying a citation." Verifying a citation and getting help with research suggest more serious purposes, but finding a book may include finding a work of fiction for reading pleasure during leisure time. Thus this source was classified as having mixed purposes.</p>

Source (and specific forum for forum sources)	Topic	Purpose
World of Warcraft Chat	<p>Gaming</p> <p>World of Warcraft is a game, thus it was assumed that the majority of conversation would focus on the game itself.</p>	<p>Recreational</p> <p>At the time of data collection, this page: http://us.blizzard.com/en-us/company/about/ described Blizzard Entertainment®, the maker of World of Warcraft, as “a premier developer and publisher of entertainment software.” The page went on to explain that “by focusing on creating well-designed, highly enjoyable entertainment experiences, Blizzard Entertainment has maintained an unparalleled reputation for quality since its inception.”</p> <p>On http://us.blizzard.com/en-us/company/about/mission.html, Blizzard was further described as “dedicated to creating the most epic entertainment experiences...ever.”</p> <p>This suggested that the intended purpose of World of Warcraft was to provide entertainment, thus recreation.</p>
AOL Chat	<p>Other, Recreational</p> <p>No descriptions of the rooms were given on AOL’s website. Rooms were most often named using a place theme, which also did not allude to any particular topic. For example, rooms were often named after public meeting places such as bars, benches, houses, etc. The choice of generalized public meeting places for names did suggest, somewhat, that the purpose was recreational in nature. In other words, no rooms were named “The Office” or “The School Quad.” Furthermore, Paolillo and Zelenkauskaitė (2013, p. 114) claim AOL chat can be “considered to have been designed for recreational, rather than serious, uses.”</p>	
NPS Chat Corpus	<p>Other, Recreational</p> <p>There is nothing on the corpus’ website or in any of the articles posted about the corpus that explain purpose or topic of the chatrooms sampled. It does say that age-specific rooms were sampled. The conversation in these rooms tends to focus on Other topics for non-serious purposes.</p>	

Appendix D: Coding Rules

This list was used by the researcher and the external coder to classify terms by linguistic feature. For each feature, a basic definition is provided with an example(s) that helped guide the classification of terms by feature. Exceptions to classifications are noted with the “Don’t count” header to indicate instances where a term would not have been classified by that feature. The “If, Also Mark As” header lists cases where a term should be classified by more than one feature. Some features also have a “Notes” header for other general information that should have been considered during classification. What appears below is what both the researcher and the coder used to classify terms by feature.

Some terms you will classify will seem like ordinary English to you. For example, I have kept the acronym *DVD* (*digital video disc*) in the list of terms I'm analyzing (instead of throwing it out as general/standard English). I will agree with you that this is rather ordinary and not a true refashioning of general English. However, for the time being I am classifying *all* acronyms, even the common ones. Later, my plans are to separate acronyms into common and uncommon acronyms. However for the purposes of this test, sifting out common acronyms from uncommon acronyms is unnecessary. I have made other similar decisions about other classification categories, such as shortenings.

The following is the list of classification categories or cyberlanguage features. You have seen many of these features used in general English. The difference with cyberlanguage is that these features are thought to be used more heavily in online communication than in ordinary writing.

For each feature, you will be presented with its feature name and a general definition. Some features will include additional information such as when NOT to count a term as this feature and when to count a term as another feature in addition to the feature in question.

Notes:

A proper noun / named entity includes: locations (states, countries, WoW zones, WoW continents, WoW battlegrounds, WoW dungeons, etc.), products and named services (sold online, bought in the store, etc.; e.g., Kleenex, Dropbox), titles (of songs, books, albums, movies, etc.), etc. It doesn't include things like courses (e.g., algebra, chemistry) or roles (e.g., assistant professor).

SINGLE LETTER FORMS (SLF)

Definition: A single letter is used in place of an entire word (e.g., *O* for *offense*). The word must contain the letter. Usually the letter is the first letter in the word. This assumes some sort of Shortening has taken place, but do not also count as a Shortening.

- Some single letter forms are not letter homophones (e.g., *D* for *defense* and *O* for *offense*).

Don't Count: Use of initials in citations/references (e.g., *Smith, A.* rest of citation...) are not counted. Or if someone uses a person's full first name and their last name initial, e.g., *Mike B.* – too ordinary, don't count the last name initial as a Single Letter Form. Also, in the SMS corpus, it is important to be aware that Dr. Sotillo deidentified this corpus by providing only a single letter as a stand-in for a person's name or a last initial. These should not be counted as Single Letter Forms.

If, Also Mark As: It doesn't matter if these are capitalized or not unless the word is a person's name or a proper noun or part of an all caps utterance.

- If it is a person's name or proper noun and is not capitalized, then mark also as Lowercase.
- If the Single Letter Form is a part of an all caps statement, then count as All Caps.

LETTER HOMOPHONES (LH)

Definition: A single letter is used in place of a sound (e.g., *U* for *you*, *r* for *are*). In other words, the sound when pronouncing the letter mimics the part of the word or whole word it is replacing. This assumes some sort of Shortening has taken place, but do not also count as a Shortening. Some letter homophones will also be Single Letter Forms: e.g., *u* for *you* is a Letter Homophone and a Single Letter Form, as is *r* for *are*.

If, Also Mark As: If the sound is the entire word. If so, then also mark as Single Letter Form.

NUMBER HOMOPHONES (NH)

Definition: A single numeral is used in place of a sound. In other words, the sound of the numeral mimics the part of the word or whole word it is replacing (e.g., *8* sounds like *ate*).

ACRONYMS (ACRO)

Definition: The initial letters for the words in a phrase or syllables within a word (e.g., *HoL* for *Halls of Lightning* and *MgT* for *Magister's Terrace*) are used in place of the full phrase. It is possible that instead of an initial letter, an interlocutor uses a number homophone – e.g., *G2G* for *good to go*, or the interlocutor may use a number standing in for a word in the acronymized phrase – e.g., *LF1M* for *looking for one more*. These should also be counted as acronyms.

Don't Count: Don't count *e.g.* or *i.e.* as acronyms – too common, more often used as the acronym than spelled out. However, if they don't include the punctuation (e.g., *ie*), then mark as Punctuation Omission.

If, Also Mark As:

- An acronym should be capitalized if:

- If it represents a proper noun or person's name (i.e. named entity). If it is not capitalized, mark as Lowercase.
 - E.g. *asist* – *American Society for Information Science & Technology*: should be marked as Lowercase
- If the acronym stands for a proper noun that includes articles or prepositions, the letters standing in for the articles and prepositions may be lowercase.
 - E.g. *CoT* – *Caverns of Time*: doesn't need to be marked as Lowercase.
- The letters for an in-word syllable may also be lowercase.
 - E.g. *MgT* – *Magister's Terrace*: doesn't need to be marked as Lowercase.
- If the acronym includes the pronoun *I* as in *idc* for *I don't care*, mark as Lowercase because the *I*, at least, should be capitalized.
- If the acronym is for a degree earned: *PhD*, *MS*, etc., it should follow standard capitalization conventions. If it doesn't mark as Lowercase (e.g. *phd*).
- An acronym with the word *god* in it doesn't require capitalization for this corpus at this time (e.g., *omg* for *oh my god* is fine in lowercase and doesn't need to be marked as Lowercase).
- Acronyms for ordinary phrases are acceptable in lowercase:
 - *btw* for *by the way*
 - *lol* for *laughing out loud*

Notes:

- *SC1* for *StarCraft 1* – the *SC* is the acronym, the *1* is really not a part of the acronym (for the phrase or named entity name) in this case. It's a version number acting as a modifier for the acronym. So this type of thing should be classified as an Acronym with Space Omission. Numbers meant as numbers (e.g., version numbers, numbers of people for a dungeon/instance) rather than standing in for a word in the acronymized phrase/named entity should not be considered part of the acronym. For example *ToC10* – the dungeon/instance is called *Trial of the Crusader*, not *Trial of the Crusader 10*. You can do it with 10 people or with 5 people. That's a number, not a true part of the name of the dungeon/instance, it acts almost like a version of ToC in this case.

STATE ABBREVIATIONS (SA)

Definition: Abbreviations found on this page: <http://www.stateabbreviations.us/>

Don't count: Don't count also as an Acronym. If the standard abbreviation form doesn't include a period (as indicated on the above webpage), just leave it alone –don't count as Punctuation Omission.

If, Also Mark As:

- All state abbreviations should be capitalized (with standard the first letter, and with postal both letters). If it is not capitalized in the way the above webpage capitalizes them, then also mark as Lowercase.

- If the standard abbreviation form doesn't include a period (as indicated on the above webpage), then also count as Punctuation Omission.

CLIPPINGS (CLI)

Definition: The final character (or digit) has been removed (e.g., *movin* for *moving*).

Don't Count: If more than one of the last letters has been removed (i.e. an uninterrupted series of final letters), don't count this as a Clipping. Instead count as Shortening.

- E.g., Ordinarily *fuckn* would be counted as a Clipping and Shortening, however, the rules for Phonetic Respelling indicate that it should be counted as Phonetic Respelling instead.
- E.g., *lev* for *level* is a Shortening not a Clipping because both the *e* and *l* were omitted which constitutes the last syllable. If only the *l* were omitted, this would be a Clipping.

If, Also Mark As: If a single quote (') has been inserted in its place, then count that also as a Phonetic Respelling.

SHORTENINGS

Definition:

- The removal of multi-letter syllables or the removal of individual letters from a word for cases that do not fall under Clipping, Single Letter Form, Letter Homophone, or Acronym. This applies only to full words, not missing letters from words like acronyms.
 - E.g., *palis* for *pallies*. The affixation (adding a *-y*) is pluralized. So a *y* should change into an *ie* when made plural. The removal of the *e* signifies a Shortening. Because an extra *l* would be added to turn *paladin* into *pally*, the missing *l* also connotes a Shortening.
- Phrase shortening: when an entire word is removed from a multi-word term, then count that as a shortening as well. E.g., *nakies* for *naked pictures*.

Don't Count:

- Don't count any variation (uppercase, lowercase, with periods, without periods) of *ok*, such as *okay*, *OK*, *O.K.*, or *o.k.*. However, *kk* or its variants is a viable cyberlanguage candidate.
- Regarding shortenings such as *'04* or *04* for *2004*, don't count. Do not count as Symbolic Substitution either.
- If a letter is omitted from an Acronym, this should not be counted as a Shortening (count as a Misspelling/Typo instead).

If, Also Mark As:

- Shortenings, particularly removed vowels, may be a part of a Phonetic Respelling. They should be marked as Phonetic Respelling only if the reduction's goal is to

simulate prosodic effect. In other words, if the interlocutor spells the word by omitting letters to indicate that s/he is choosing not to pronounce those letters, then it is Phonetic Respelling only (e.g., *cud* for *could*)

- If the Shortening doesn't add or mirror prosody/pronunciation, then it should be marked only as Shortening. In the majority of cases, a Phonetic Respelling will not also be classified as a Shortening.
- Examples where a Phonetic Respelling should also be marked as Shortening:
 - *thnxx* (the *a* sound is pronounced)
 - *gnna* (the *o* sound is pronounced)

PUNCTUATION OMISSION (PO)

Definition: Omission of punctuation within a word (but not sentence ending punctuation such as periods, question marks, or exclamation points that should appear at the end of a sentence). E.g., apostrophes missing in contractions like *dont* for *don't*.

- e.g. and *i.e.* should include the periods. If they don't, classify as Punctuation Omission (but not Acronym).

Don't Count:

- Merriam-Webster and Oxford may say that a term should include a dash (e.g. *e-journal*, *dum-dum*). If the dash is missing, don't count as Punctuation Omission. There are probably other dictionaries that say the lack of dash is okay.
- Regarding elisions or other phonetic respellings such as *ima* and *imma* and *dunn*, don't count as Punctuation Omission. Count as Phonetic Respelling only.

If, Also Mark As:

- If the pronoun *I*, in lowercase, is used in an elision (e.g., *ima*, *imma*), then mark also as Lowercase.

SYMBOLIC SUBSTITUTION (SYM SUB)

Definition: A letter, punctuation mark, and rarely a numeral is substituted for an entire word (multi-character word) or larger idea/concept that doesn't contain that letter or digit. (This should not overlap with Single Letter Forms, Letter Homophones, or Number Homophones.) Most of the time it is a punctuation mark (instead of a letter) that is being used to replace the word. E.g., *apples* > *bananas* to symbolize that apples are better than bananas.

- This can also include using letters or punctuation marks to write a curse word in a way that attempts to disguise the fact that it is a curse word (e.g., *f*%! for fuck*). For example, for the purpose of trying to be more polite. Or it can include ways to deidentify a named entity (e.g., instead of *John*, *J---*). This form of deidentification is not something the corpus author (the researcher or those from whom the researcher has borrowed corpora) added. It is something the interlocutor included.
- This can also include using numbers or punctuation marks in place of letters whose orthography is similar to the numbers/punctuation marks. E.g. *1337* for *leet*. As you can see with this example, none of the numerals are Number Homophones.

- Also, a punctuation mark used to signify something being done to the word itself. E.g., the slash in *b/c* is signifying acronymy.
- Could include using *x* as a “fill in the blank here” type of situation: E.g., *Mr. X, x number of people*
 - Similar to *x* in this regard, any full word that is standing in for a missing word or concept. E.g., *[whatever] Books* – the person can't remember the name of the first part of the bookstore name, only that the second part is *Books*, as in *Quail Ridge Books*.

Don't Count:

- Exceptions are noted on the Table of Signs and Symbols.

CONJUNCTIONS AND DISJUNCTIONS

Definition: A slash is used in place of *and* or *or* to symbolize that the concepts joined by the slash exhibit an AND or OR condition.

- E.g., *peanut butter/jelly sandwich* for *peanut butter and jelly sandwich*
- E.g., *which one wants to do that? you/joe?* for *which one wants to do that? you or joe?*

Also Mark As: Conjunctions and disjunctions are special cases of Symbolic Substitution, so also mark as Symbolic Substitution.

ALL CAPS (CAPS)

Definition: An entire word in caps (e.g., *NOOO, EXACTLY*). Other instances:

- a capitalized Single Letter Form (or Letter Homophone) in an all caps utterance (e.g., *B RIGHT BACK!*)

Don't Count:

- *A* if it is the first word in a sentence (as in *A dog walked by.*)
- *I* (first person pronoun)

However, if the entire sentence is in caps, then an *I* and *A* should be counted as all caps (e.g., *A BIG BIRD FLEW BY!* or *I AM NOT HAPPY!*)

LETTER DUPLICATION (L DUP)

Definition: The same letter is used 2 times or more than what is required to spell the word correctly.

- E.g., *look* – not Letter Duplication
- E.g., *looooooook* – yes Letter duplication
- E.g., *hi* – not Letter Duplication
- E.g., *hii* – yes Letter Duplication

Don't Count: While this is sometimes a form of Phonetic Respelling, don't count also as Phonetic Respelling. It is possible that the interlocutor didn't mean to type so many

letters and if that is the case, double-classifying it as Phonetic Respelling also would be an error.

PUNCTUATION DUPLICATION (P DUP)

Definitions: Two or more punctuation marks in succession.

- E.g. ?! yes Punctuation Duplication
- E.g. ~~ yes Punctuation Duplication
- E.g. *no way....* yes Punctuation Duplication
- E.g. "happy" not Punctuation Duplication
- Punctuation Duplication can be used in Offsetting (e.g., <<<<<JOE>>>>>).
- Punctuation Duplication includes repeated units of punctuation as in:
 - o *(*(*Harry*))* – two repeated units of *(** and **)*

Don't Count:

- However emoticons (which can be considered Punctuation Duplication of a sort) should not be counted as Punctuation Duplication.
- Punctuation Duplication doesn't include ellipses (of any length) that are used to suggest an omitted part of a quoted phrase. In other words, ellipses used in a standard way in the middle of a quote (e.g., *Abraham Lincoln said, "Four score and twenty years ago, our fathers brought forth on this continent, ... in Liberty and dedicated to the proposition that all men are created equal."*) should not be counted as Punctuation Duplication. In other words, standard use of ellipses when referencing a quote from an author and needing to indicate that part of the quote has been omitted should not be counted as Punctuation Duplication.

If, Also Mark As:

- If the duplicated punctuation is joined by 2 or more words (or an emoticon), also mark it as a Compound/Space Omission.
 - o E.g., *thanks...^^* mark also as Compound/Space Omission
 - o E.g., *happy...not* mark also as Compound/Space Omission
 - o E.g. *happy...* do not mark as a Compound/Space Omission
- If the duplicated punctuation is being used to offset, then also mark as Offsetting (e.g., <<<<<COOL>>>>>)
- Mark also as Symbolic Substitution if a series of periods are being used to stand in for *etc.* Usually when this happens, the periods are preceded by a comma-delimited list of words (e.g., *at the fair, i ate popcorn, cotton candy, funnel cake,*).

PHONETIC RESPELLING (PR)

Definition: Changing the spelling of a word to mimic the prosodic effects or pronunciation of the word.

- Phonetic Respelling includes elisions (combining 2 or more words together in a way to indicate the prosodic effect when a speaker verbally runs the words together). These should not be counted also as Space Omission.
 - E.g., *gonna, gotta, whatcha, coulda*
- Phonetic Respelling includes substituting letters for the correctly spelled letters to indicate prosodic effect, including accent.
 - E.g., *sucka, nuthin, thx, tunez*
- *lil* for *little* will be counted as a Phonetic Respelling instead of as a Shortening at this time.
- This can include situations which are, to some extent, the reverse: e.g., using *ph* instead of an *f* to draw attention to the sound. E.g., *phail* for *fail*.

Don't Count:

- The removal of one consonant in a double medial consonant pattern. Instead count as a Shortening. E.g. *formaly* for *formally* should be counted as a Shortening.
- This includes two consonants that have the same sound: e.g., *truck* – if the *c* were removed, this should be counted as a Shortening, not Phonetic Respelling.
- The silent *e* that some people omit in words that end in *-ly* is often not Phonetic Respelling (e.g., *immediatly* – count as a Shortening instead).

If, Also Mark As:

- If a Clipping uses ' in place of the clipped letter, count also as Phonetic Respelling.
- If the pronoun *I*, in lowercase, is used in an elision (e.g., *ima, imma*), then mark also as Lowercase.
- If the Phonetic Respelling is for a proper noun and it is not capitalized, mark also as Lowercase.

Notes

- Phonetic Respellings may include other features such as Shortenings. They should be marked as Phonetic Respelling only if the shortening's goal is to simulate prosodic effect. In other words, if the person omits letters in a way to signify that s/he is not pronouncing them, then it is Phonetic Respelling only (e.g., *cud* for *could*, *n* for *and*). This also includes the removal of silent consonants such as *gh* in *strait* for *straight*. (This doesn't include the removal of a single silent *e* in *-ly* affixes however.)
 - If the reduction doesn't add or mirror prosody/pronunciation, then it should be marked only as a Shortening. In the majority of cases, a Phonetic Respelling will not also be classified as a Shortening.
 - The only examples I found where it should also be marked as a Shortening:
 - *thnxx* (the *a* sound is pronounced) or similar variants that leave out the *a*
 - *gna* (the *o* sound is pronounced)

SPELLING ALOUD (SP A)

Definition: When someone specifically spells or pronounces letters within the word instead of the saying the entire word. E.g., *Y O U*.

ONOMATOPOEIC EXPRESSION (ONO)

Definition: Sounds of various kinds. This includes:

- Minimal responses and other human vocalizations:
 - o Minimal responses: *hmm, uh, oh, huh, erm*
 - o Human vocalizations: *haha, hehe, grr*
 - o Exclamations/interjections: *woot, whoa, awww, pfft, ah, aha!*
- Sound effects in the environment: *bam, crasssh, splaaat*
- Singing sounds/notes (not actual words): *da da dum dum doooo*

OFFSETTING PUNCTUATION (OP)

Definition:

- Punctuation used to emphasize a word or phrase, placed **on both sides** of the word or phrase (in place of bolding or italicizing which is usually unavailable in these media). E.g., ****Happy Birthday****
- Punctuation (but not quotes) used to demarcate an emote (on both sides of the emote word or phrase). E.g., ***kicks Lou in the butt***
- Punctuation used to decorate or draw attention to a word/phrase in an utterance so it will visually stand out. E.g., **wb (*((*JOE*) *) *)** or **<Kung Fu Pandas> is recruiting all classes**
- Use of asterisks (or other non-traditional punctuation) instead of quotation marks to indicate titles of things (e.g., song titles, album titles, book titles, etc.) or quoted phrases. E.g., **the song *Ave Maria* is one of my favorites**
- Also asterisks in place of parentheses.

If, Also Mark As: If the punctuation on either end of the offset word/phrase is duplicated, also mark as Punctuation Duplication. E.g., **(*((*JOE*) *) *)** or **<<<<JOE>>>>**

EMOTICONS

Definition: Using a series of letters, punctuation, and/or numbers to create a pictogram of a face. E.g., **^^**

EMOTES

Definition: Text meant to indicate:

- The interlocutor's actions (e.g., *burp, cough, smiles, looks for his poking stick*)
- The state of being of the interlocutor—emotional, physical, personality characteristic (e.g. *is a smoker, happy, chocolate grin*)

Don't Count: Laughter (e.g., *haha*, *hehe*, etc.) should not be counted also as an Emote.

If, Also Mark As:

- Emotes can include Onomatopoeic Expression if they satisfy Onomatopoeic Expression conditions. The emote must be spelled in a way intended to indicate sound.
 - E.g., *burp* just an Emote
 - E.g., *burrpppp* Emote, Onomatopoeic Expression, and Letter Duplication
- If an Emote is enclosed in punctuation of some sort, it should also be classified as Offsetting (e.g. **looks for his poking stick**)

POINTING (PTING)

Definition: The use of <, >, and possible dashes (e.g., -->) to indicate an arrow of some sort to point to one's self, another person, or some other part of the utterance.

Don't Count: It should not be classified as Offsetting.

Also Mark As: If the arrow points to the interlocutor or to another interlocutor, then this is a type of Emote and should be marked as both Pointing and Emote.

ADDRESSIVITY (ADD)

Definition: Use of characters to indicate a specific person to whom one wishes to address his/her comments. E.g *@Laura*

PICTOGRAM (PICT)

Definition: Using characters to create a graphical/visual representation of a physical object

Don't Count: Technically this includes emoticons; however, double classification (as both emoticon and pictogram) isn't necessary at this time. Emoticons are a class unto themselves really.

AFFIXATION / COMBINING FORMS (AFF)

Definition:

- adding prefixes/suffixes to bases, especially unconventional bases like acronyms or shortenings, to form words not found in Merriam-Webster or Oxford (e.g., *lolers*)
- adding combining forms to bases, especially unconventional bases like acronyms or shortenings, to form words not found in Merriam-Webster or Oxford (e.g., *multi-dps*)

- adding verb tense endings to only unconventional bases like acronyms or shortenings, to form words not found in Merriam-Webster or Oxford (e.g., *dpsing*)
- adding affixes that are used as diminutives/familiarity markers (e.g., suppose your character's name is *Narn*, and gets changed into *Narnie*).

Notes:

- What about rather ordinary instances?
 - o E.g., subdomain
 - o 40ish
 Look the affix/combining form up in the dictionary to see if the entry makes allowances for the usage found in the corpus. E.g., In a dictionary, it said *-ish* can be used with ages. The OED says it can be added to hours of the day or numbers of the year. If the dictionary makes allowances for this, then don't count.
- If Merriam-Webster or Oxford say that the affix should have been appended with a dash or with no space (e.g., *co-gm*) but the dash was omitted or a space left in (e.g., *co gm*), then still count as Affixation, but don't worry about counting it also as Punctuation Omission. There may be other dictionaries or lexicographers that would say this is okay.
 - o *E* as in *e journal*, *e-journal*, *e-book*, etc. will be counted as a prefix.

BLENDS

Definition: A compounding of 2 shortened/abbreviated terms. Example: *brunch* for *breakfast* and *lunch*. Only mark as a Blend though (not also as a Shortening or Compound).

COMPOUNDS / SPACE OMISSION (COMP / SO)

Definition: Space Omission:

- Some instances are typo-ish: the space was omitted, mostly likely unintentionally (e.g., *ifthey*). This is true Space Omission.
- Many are cases of series of periods and words/emoticons on either side of the periods as in *thats why...:-)*.
- Some are cases of a number followed by a unit of measure (e.g., *3min*, *45rpm*).

Definition: Compounding: Other instances seem to be where a person has created a new word by compounding. Usually these have dashes (e.g., *drip-irrigation*, *double-staffed*). These terms are not found in Merriam-Webster or Oxford.

Don't Count:

- Some cases are numbers followed by *k* for thousand. Don't count as Space Omission because *1k = 1000* and it wouldn't be *1 k = 1 000*. There's no space in the number in other words, so no space in the abbreviation. Ultimately, *1k* wouldn't be counted at all anyway per the Table of Signs and Symbols.

- *100g* or other variants where *g* means *gold*, because the game writes it this way and it mimics *\$100*. There is no space between the *\$* and *100*.
- Don't count standard usages of a dashed compound that is being created to form an adjective, e.g., *top-grossing film*, *long-considered idea*, etc.
- If the compound (or affixed form) appears in Merriam-Webster or Oxford in any form (with a dash, without a dash but with a space, without a space or dash), don't count it. There are probably other dictionaries that say the way it was written in the corpus is okay.
- Don't count acronyms of any kind as Space Omission. Space omission is a natural part of acronymy.

Notes:

- Will separate instances into true Compounds and just Space Omission.
- *Log in* (verb), *login* (noun). If the verb sense is being used but the space has been omitted, that should be Space Omission.

CONVERSION (CONV)

Definition: From Quirk et al. (1985): “Conversion is the derivational process whereby an item is adapted or converted to a new word class without the addition of an affix” (p. 1558). This means turning an adjective into a noun, a noun into a verb, a noun into an adjective, etc. Affixes that are verb endings or a plural/possessive are okay here – it’s the true prefixes and suffixes like *-ment*, *pre-*, *-ish*, etc. that should not be a part of the word formation if it is to be considered conversion. E.g., *Door* is a noun. If someone were to change it to a verb as in “*I doored him in the face*” then that would be Conversion, not Affixation. However, *piggish* is not an example of Conversion because it is using the affix *-ish* to change *pig* to an adjective. It is Affixation.

REDUPLICATION (REDUP)

Definition: From Quirk et al. (1985): “compounds that have 2 or more constituents which are either identical or only slightly different” (p. 1579). Examples: *goody-goody*, *walkie-talkie*, *din-din*, *ha ha*, *bow woow*, *flip-flop*, *dilly-dally*, *tip-top*.

Don't Count: For this study, only count things that are exact duplicates (e.g., *goody goody*, *din-din*, *kk*) not similar terms (e.g., *flip-flop*, *tip-top*). And do not count onomatopoeic laughter (e.g., *haha*) as reduplication. If a reduplication has no spaces (e.g., *yesyesyes*), then don't count as Space Omission.

LOWERCASE (LC)

Definition: Mark as Lowercase if the word is:

- The pronoun *I* (e.g., *i* should be marked as Lowercase)
- Proper name/named entity (either in full form or as a shortening, acronym, single letter form). E.g., mark *harris teeter* or *civic* (for a Honda Civic) as Lowercase.
 - These do not include:

- The names of school subjects (e.g., algebra, calculus)
- The names of drugs (e.g., low dose naproxen, aspirin)
- Titles: e.g., *Dr, Mr, Mrs* (e.g., *dr* should be marked as Lowercase)
- Street abbreviations if referencing a specific street: e.g., *Rd, Ave* (e.g., both the *franklin* and the *st* in *franklin st* should be marked as Lowercase)
- See the Acronyms section for specific guidance on capitalization of acronyms.
- Words that are in all caps (e.g. NVIDIA) but the interlocutor only capitalized the first letter (or words that are intended to have certain letters capitalized that the interlocutor made lowercase) should be marked as Lowercase.

Notes: A proper noun / named entity includes: locations (states, countries, WoW zones, WoW continents, WoW battlegrounds, WoW dungeons, etc.), products and named services (sold online, bought in the store, etc.; e.g., Kleenex, Dropbox), titles (of songs, books, albums, movies, etc.), etc. It doesn't include things like courses (e.g., algebra, chemistry) or roles (e.g., assistant professor).

MISSPELLING/TYPO (MT)

Definition:

- An extra space that shouldn't be there (e.g., *I dont want to do that !*) – almost the opposite of Space Omission.
- An extra character(s) that makes the word incorrectly spelled (e.g., *filnd* for *find*).
 - This includes punctuation marks that are not a part of a named entity. E.g., *e-bay* for *eBay* is a misspelling.
- An extra character (in particular, a punctuation mark) that doesn't fit (e.g., like ending an utterance with *@!* instead of *!!*).
 - Note that question marks and exclamation marks combined should not be counted as a Typo (e.g., *really??!* should only be counted as Punctuation Duplication).
 - Note that a series of periods followed by a single question mark or exclamation mark should not be counted as a Typo (e.g., *i dunno...?* should not be counted as Misspelling/Typo).
- Letters transposed (e.g., *laern* for *learn*).
- An incorrect letter in place of the correct one (e.g., *dammed* for *damned*, or *lom* for *lol*)
- An ordinary word that has a letter capitalized after the first letter (e.g., *LEft*)
- A letter missing from an acronym. The letter should represent a noun, verb, adjective (e.g., *dp* for *damage per second*) not an article or preposition. If the letter represents an article or preposition, then don't count as a typo (e.g., *LG* for *looking for group*).

Don't Count:

- What to do with things that contain dashes or spaces when Merriam-Webster says there shouldn't be dashes/spaces (or the reverse situation):

- Some are clear Misspelling/Typos: *base ball* for *baseball* – *baseball* is always spelled this way in every context and dictionary, so *base ball* should be counted as a Misspelling/Typo.
- Some aren't as clear: *co gm* instead of *co-gm* or M-W says *coproducer* but the person wrote *co-producer*.
 - o Don't count standard usages of a dashed compound when using ordinary adjectives, e.g., *top-grossing film*, *long-considered idea*, etc.
 - o If the compound appears in Merriam-Webster or Oxford in any form (with a dash, without a dash but with a space, without a space or dash), don't count it UNLESS it is always the way Merriam-Webster and Oxford have it – all dictionaries/writing (e.g., *baseball*).
 - o Because *co gm* or *co-gm* aren't in a dictionary, count as affixation and don't count a missing dash as a typo.
- If *-ize* is spelled with an *s* instead of a *z* (e.g., *intellectualise*), don't count as anything. This is okay.

REPAIRS

Definition: Indication of repairing a typo/misspelling, a disfluency (e.g., **the* for *the*). This may include some sort of punctuation, but in most cases, don't count a single punctuation mark on one or both sides of the repaired term as anything (e.g., **stop** repairing the omission of *stop* in the utterance “*i wanted to doing that*” should not be counted as anything other than a Repair; don't count as Offsetting in other words).

FORMATTING WORKAROUNDS

Definition: Because these media often do not allow for things like bulleting, indentation, etc., any attempt to achieve this via other means in a Formatting Workaround. This also includes using asterisks to indicate possible answers to a question because this is a way to do a bulleted list in a sense. Also, if a dash is used to sign a message, count as Formatting Workarounds. E.g. Special ways to do footnotes: *[*]* and *[I]*

Don't count: Dashes used as bullets, because Microsoft Word has dashed bullets.

Appendix E: Table of Signs and Symbols

What appears below is an addendum to the Coding Rules, used by both the researcher and the coder to classify terms by feature. The table below lists signs and symbols that are sometimes used in conventional ways. When used conventionally, instances of these signs and symbols were not classified as symbolic substitution. These conventional uses are outlined in the Meaning column. However, when used in unconventional ways, they were classified as symbolic substitution. These exceptions—i.e., unconventional ways—are noted below in bold.

In most cases, when you see one of the signs/symbols listed below, you will not count these as Symbolic Substitution (SS). However, there are cases when you would count them as Symbolic Substitution; notes are included in the Meaning column when this is desired. These exceptions are noted in red.

Sign/Symbol	Meaning
%	percent, per hundred (e.g., 25%)
+	plus, addition (e.g., 100+), with numbers only positive
-	minus, subtract (e.g., 10-5 = 5) negative (e.g., -5)
-	range (e.g., 20-25%)
&	and (e.g., cats & dogs)
@	at (e.g., call me @ home) However, if it is used in place of an <i>a</i>, count as Symbolic Substitution: c@ke for cake.
\$	dollar sign—used universally for monetary units (e.g., \$5) However, if the sign is repeated, then count as Punctuation Duplication.
=	equality, equals, equal to (e.g., 4+5 = 9,) with numbers only However if it is used with words, then count as Symbolic Substitution: e.g., pizza = yum
/	divided by (e.g., 4/2)
/	per (e.g., \$6/lb), with prices or costs only However if it is used with something other than a price (e.g., 30 students/class for 30 students per class) then count as Symbolic Substitution.
<	less than (e.g., 4 < 5), with numbers only However with words as in bananas < applies, count as Symbolic Substitution.

Sign/Symbol	Meaning
>	greater than (e.g., $5 > 4$), with numbers only However with words (e.g., <i>tacos > pizza</i>), count as Symbolic Substitution.
#	number, numbered (e.g., #4) However used as decoration of some kind (e.g., ##COOL###) count as Offsetting.
#	pound (e.g., 2# <i>watermelon</i>), for weights only
~	equivalent to, similar to
~	approximately (e.g., ~400)
x	multiplied by (e.g., $4 \times 5 = 20$), with numbers only However used with an object, as in <i>cookiesX20</i>, count as Symbolic Substitution.
x	by (for dimensions) (e.g., 4"x4")
'	foot, feet (e.g., 5'x6')
'	to symbolize the omission of the first two digits in a year (e.g. '09 for 2009)
"	inches (e.g., 4"x4"), or ditto
*	multiplied by (e.g., $4 * 5 = 9$)
k (K)	1000
Rx	take (as in prescription)
AM	ante meridian
PM	post meridian
1 st , 2 nd , 3 rd , etc.	abbreviations for numerals

References

- Adams, M. (2009). *Slang: The people's poetry*. Oxford; New York: Oxford University Press.
- Albee, E. (1959). *The American Dream and The Zoo Story: Two plays by Edward Albee*. New York: Signet.
- Androutsopoulos, J. (2006). Introduction: Sociolinguistics and computer-mediated communication. *Journal of Sociolinguistics*, 10(4), 419–438. doi:10.1111/j.1467-9841.2006.00286.x
- Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596.
- Atkins, B. T. S., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford; New York: Oxford University Press.
- Ball, C. N. (1994). Automated text analysis: Cautionary tales. *Literary and Linguistic Computing*, 9(4), 295–302. doi:10.1093/lc/9.4.295
- Baron, N. S. (2003). The language of the Internet. In A. A. S. Farghaly (Ed.), *Handbook for language engineers* (pp. 59–127). Stanford: CSLI Publications.
- Baron, N. S. (2008). *Always on: Language in an online and mobile world*. Oxford: Oxford University Press.
- Baron, N. S. (2010a). Are instant messages speech? In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International handbook of Internet research* (pp. 1–21). Springer Netherlands.
- Baron, N. S. (2010b). Discourse structures in instant messaging: The case of utterance breaks. *Language@Internet*, 7. Retrieved from <http://www.languageatinternet.org/articles/2010/2651>
- Baron, N. S. (2010). Are instant messages speech? In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International handbook of Internet research* (pp. 1–21). Springer Netherlands. Retrieved from http://link.springer.com.libproxy.lib.unc.edu/chapter/10.1007/978-1-4020-9789-8_1

- Baron, N. S. (2013). Instant messaging. In S. Herring, D. Stein, & T. Virtanen (Eds.), *Pragmatics of computer-mediated communication* (pp. 135–161). Berlin: De Gruyter Mouton.
- Baron, N. S., & Ling, R. (2011). Necessary smileys and useless periods: Redefining punctuation in electronically-mediated communication. *Visible Language*, 45(1), 45–67. Retrieved from <http://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ933707>
- Bauer, L. (1983). *English word-formation*. Cambridge: Cambridge University Press.
- Beghtol, C. (2001). The concept of genre and its characteristics. *Bulletin of the American Society for Information Science*, 27(2), 17–19. Retrieved from <http://www.asis.org.libproxy.lib.unc.edu/Bulletin/Dec-01/beghtol.html>
- Berkenkotter, C., & Huckin, T. N. (1995). Rethinking genre from a sociocognitive perspective. In *Genre knowledge in disciplinary communication: Cognition/culture/power* (pp. 1–26). Hillsdale, New Jersey: L. Erlbaum Associates.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. (2008). Representativeness in corpus design. In T. Fontenelle (Ed.), *Practical lexicography: a reader* (pp. 63–87). Oxford; New York: Oxford University Press.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press. Retrieved from <http://search.lib.unc.edu/search?R=UNCb3145930>
- Bieswanger, M. (2007). 2 abbrevi8 or not 2 abbrevi8: A contrastive analysis of different space and time-saving strategies in English and German text messages. *Texas Linguistics Forum*, 50. Retrieved from <http://studentorgs.utexas.edu/salsa/proceedings/2006/Bieswanger.pdf>
- Bieswanger, M. (2013). Micro-linguistic structural features of computer-mediated communication. In S. Herring, D. Stein, & T. Virtanen (Eds.), *Pragmatics of computer-mediated communication* (pp. 463–485). Berlin: De Gruyter Mouton.
- Blizzard Entertainment. (2010, October 7). *World of Warcraft® subscriber base reaches 12 million worldwide*. Retrieved May 11, 2012, from <http://us.blizzard.com/en-us/company/press/pressreleases.html?id=2847881>

- Boneva, B., Quinn, A., Kraut, R., Kiesler, S., & Shklovski, I. (2008). Teenage communication in the instant messaging era. In R. Kraut (Ed.), *Computers, phones, and the Internet: Domesticating information technology* (pp. 201–218). Oxford: Oxford University Press.
- Brachman, J. M. (2006). High-tech terror: Al-Qaeda's use of new technology. *Fletcher Forum of World Affairs*, 30(2), 149–164.
- Cabré, M. T. (Maria T. (1999). *Terminology: theory, methods, and applications*. (J. C. Sager, Ed.). Amsterdam; Philadelphia: J. Benjamins Publishing Company.
- Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J. C., & Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1), 13–31. Retrieved from <http://dl.acm.org.libproxy.lib.unc.edu/citation.cfm?id=972684.972686>
- Carlson, J. R., & Zmud, R. W. (1999). Channel expansion theory and the experiential nature of media richness perceptions. *The Academy of Management Journal*, 42(2), 153–170.
- Carter, R. (2004). *Language and creativity: The art of common talk*. London: Routledge.
- Cherny, L. (1999). *Conversation and community: Chat in a virtual world*. Stanford, California: CSLI Publications.
- Cho, T. (2010). Linguistic features of electronic mail in the workplace: A comparison with memoranda. *Language@Internet*, 7(3). Retrieved from <http://www.languageatinternet.org/articles/2010/2728>
- Chomsky, N. (1966). Creative aspect of language use. In *Cartesian linguistics: a chapter in the history of rationalist thought* (pp. 51–71). New York: Harper & Row.
- Christopherson, L. (2013). Throwing yourself into World of Warcraft® chat with +20 haste. *Language@Internet*.
- CNN.com. (2006, March 16). 27 charged in child porn sting. *CNN.com: Law Center*. Retrieved from <http://www.cnn.com/2006/LAW/03/15/childporn.arrests/index.html?iref=allsearch>

- Collister, L. B. (2008, April 29). *Virtual discourse structure: An analysis of conversation in World of Warcraft* (Master's Thesis). University of Pittsburgh, Pennsylvania. Retrieved from <http://etd.library.pitt.edu.libproxy.lib.unc.edu/ETD/available/etd-06022008-142543/>
- Crystal, D. (1996). Language play and linguistic intervention. *Child language teaching and therapy*, 12(3), 328–344.
- Crystal, D. (1998). *Language play*. Chicago: The University of Chicago Press.
- Crystal, D. (2006). *Language and the Internet*. Cambridge: Cambridge University Press.
- Crystal, D. (2008a). *A dictionary of linguistics and phonetics* (6th ed.). Malden, MA: Blackwell.
- Crystal, D. (2008b). *Txting: the gr8 db8*. New York: Oxford University Press.
- Daft, R. L. & Lengel, R. H. (1984). Information richness: A new approach to managerial behavior and organizational design. In B. M. Staw & L. L. Cummings (Eds.) *Research in organizational behavior* (pp. 191-233). Greenwich, CT: JAI Press.
- Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science*, 32(5), 554–571.
- Danet, B. (2001). *Cyberpl@y: Communicating online*. Oxford: Berg Publishers.
- Danet, Brenda, Ruedenberg-Wright, L., & Rosenbaum-Tamari, Y. (1997). “HMMM...WHERE’S THAT SMOKE COMING FROM?” Writing, play and performance on Internet Relay Chat. *Journal of Computer-Mediated Communication*, 2(4). Retrieved from <http://jcmc.indiana.edu/vol2/issue4/danet.html>
- Davis, B. H., & Brewer, J. P. (1997). A first look at electronic discourse. In *Electronic discourse: Linguistic individuals in virtual space* (pp. 1–19). Albany, New York: State University of New York Press.
- Dresner, E. (2005). The topology of auditory and visual perception, linguistic communication, and interactive written discourse. *Language@Internet*, 2. Retrieved from http://www.languageatinternet.de/articles/2005/161/index_html

- Dresner, Eli, & Herring, S. C. (2010). Functions of the nonverbal in CMC: Emoticons and illocutionary force. *Communication Theory*, 20(3), 249–268. doi:10.1111/j.1468-2885.2010.01362.x
- Driscoll, D. (2002). The ubercool morphology of Internet gamers: A linguistic analysis. *Undergraduate Research Journal for the Human Sciences*, 1. Retrieved from <http://www.kon.org/urc/driscoll.html>
- Dürscheid, C. (2004). Netzsprache – ein neuer mythos. In M. Beißwenger, L. Hoffmann, & A. Storrer (Eds.), *Internetbasierte kommunikation*. Duisburg: Gilles & Francke.
- Dürscheid, C., & Frehner, C. (2013). Email communication. In S. Herring, D. Stein, & T. Virtanen (Eds.), *Pragmatics of computer-mediated communication* (pp. 35–54). Berlin: De Gruyter Mouton.
- Eble, C. C. (1996). *Slang and sociability: In-group language among college students*. Chapel Hill, North Carolina: University of North Carolina Press.
- Ferguson, C. A. (1994). Dialect, register, and genre: Working assumptions about conventionalization. In Douglas Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register*. Oxford University Press.
- Ferrara, K., Brunner, H., & Whittemore, G. (1991). Interactive written discourse as an emergent register. *Written Communication*, 8(1), 8–34.
- Gains, J. (1999). Electronic mail—a new style of communication or just a new medium?: An investigation into the text features of e-mail. *English for Specific Purposes*, 18(1), 81–101.
- Giles, H., Coupland, J., & Coupland, N. (1991). Contexts of accommodation: developments in applied sociolinguistics In *Accommodation theory: Communication, context, and consequence* (pp. 1-68). Cambridge: Cambridge University Press.
- Giltrow, J. (2013). Genre and computer-mediated communication. In S. Herring, D. Stein, & T. Virtanen (Eds.), *Pragmatics of computer-mediated communication* (pp. 717–737). Berlin: De Gruyter Mouton.
- Goffman, E. (1981). *Forms of talk*. Philadelphia: University of Pennsylvania Press.
- Goffman, E. (1990). *The presentation of self in everyday life*. N.Y.: Doubleday.

- Gries, S. T. (2009). What is corpus linguistics? *Language and Linguistics Compass*, 3(5), 1225–1241. doi:10.1111/j.1749-818X.2009.00149.x
- Grieve-Smith, A. B. (2007). The envelope of variation in multidimensional register and genre analysis. In E. Fitzpatrick (Ed.), *Corpus linguistics beyond the word: corpus research from phrase to discourse* (pp. 21–42). Amsterdam: Rodopi.
- Grinter, R. E., & Eldridge, M. (2001). y do tngrs luv 2 txt msg. In W. Prinz, M. Jarke, Y. Rogers, K. Schmidt, & V. Wulf (Eds.), *Proceedings of the Seventh European Conference on Computer-Supported Cooperative Work ECSCW* (Vol. 1, pp. 219–238). Netherlands: Kluwer Academic Publishers. doi:10.1007/0-306-48019-0_12
- Gross, J. (2013, April 22). Texting as a “miraculous thing”: 6 ways our generation is redefining communication. *TED Blog*. Retrieved from <http://blog.ted.com/2013/04/22/texting-as-a-miraculous-thing-6-ways-our-generation-is-redefining-communication/>
- Halliday, M. A. K. (1985). *Spoken and written language*. Victoria: Deakin University Press.
- Halliday, M. A. K. (2007a). Language and social man. In J. J. Webster (Ed.), *Language and society* (pp. 65–130). New York: Continuum.
- Halliday, M. A. K. (2007b). The users and uses of language. In J. J. Webster (Ed.), *Language and society* (pp. 5–37). New York: Continuum.
- Hård af Segerstad, Y. (2002). *Use and adaptation of written language to the conditions of computer-mediated communication* (Doctoral dissertation, Göteborg University, Sweden). Retrieved from http://www.ling.gu.se.libproxy.lib.unc.edu/~ylvah/dokument/ylva_diss.pdf
- Haythornthwaite, C., & Wellman, B. (2002). *The Internet in everyday life*. Malden, MA: Blackwell Publishers.
- Herring, S. C. (2001). Computer-mediated discourse. In D. Schiffrin, D. Tannen, & H. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 612–634). Oxford: Blackwell Publishers.
- Herring, S. C. (2002). Computer-mediated communication on the Internet. *Annual Review of Information Science and Technology*, 36, 109–168.

- Herring, S. C. (2007). A faceted classification scheme for computer-mediated discourse. *Language@Internet*, 4. Retrieved from <http://www.languageatinternet.org/articles/2007/761>
- Herring, S. C. (2012). Grammar and electronic communication. In C. Chapelle (Ed.), *Encyclopedia of applied linguistics*. Hoboken, NJ: Wiley-Blackwell.
- Herring, S., Stein, D., & Virtanen, T. (2013). Introduction to the pragmatics of computer-mediated communication. In S. Herring, D. Stein, & T. Virtanen (Eds.), *Pragmatics of computer-mediated communication* (pp. 3–32). Berlin: De Gruyter Mouton.
- Holtgraves, T. (2011). Text messaging, personality, and the social context. *Journal of Research in Personality*, 45, 92–99.
- Howden, M. S. (1984). Code and creativity in word formation. *Forum Linguisticum*, 8(3), 213–222.
- Isaacs, E., Walendowski, A., Whittaker, S., Schiano, D. J., & Kamm, C. (2002). The character, functions, and styles of instant messaging in the workplace. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, (pp. 11–20). New Orleans, Louisiana, USA: ACM. doi:10.1145/587078.587081
- Johnstone, B. (2008). *Discourse analysis* (2nd ed.). Malden, MA: Blackwell.
- Jucker, A. H., & Dürscheid, C. (2012). The linguistics of keyboard-to-screen communication: A new terminological framework. *Linguistik online*, 56(6), 39–64.
- Jung, C. (1971). *C. G. Jung: Psychological reflections; A new anthology of his writings, 1905-1961*. (J. Jacobi, Ed.). London: Routledge and Kegan Paul.
- Kadir, Z. A., Maros, M., & Hamid, B. A. (2012). Linguistic features in the online discussion forums. *International Journal of Social Science and Humanity*, 2(3), 276–281.
- Kilgarriff, A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora (pp. 231–245). *Proceedings of the 5th ACL-SIGDAT Workshop on Very Large Corpora*, Beijing and Hong Kong.
- Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics & Linguistic Theory*, 1(2), 263–275.

- Krippendorff, K. (2004). *Content analysis: an introduction to its methodology* (2nd ed.). Thousand Oaks, California: Sage.
- Labov, W. (2006). *The social stratification of English in New York City* (2nd ed.). Cambridge; New York: Cambridge University Press.
- Lave, J., & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. Cambridge: Cambridge University Press.
- Lehrer, A. (2007). Blendalicious. In J. Munat (Ed.), *Lexical creativity, texts and contexts* (pp. 115–133). Amsterdam: John Benjamins Publishing Company.
- Lewin, B., & Donner, Y. (2002). Communication in Internet message boards. *English Today*, 18(3), 21–28.
- Lindh, S. (2009). *Online computer game English: A study on the language found in World of Warcraft* (Bachelor's Thesis, Karlstad University, Sweden). Retrieved from <http://kau.diva-portal.org/smash/record.jsf?pid=diva2:221885>
- Ling, R. (2005). The sociolinguistics of SMS: An analysis of SMS use by a random sample of Norwegians. In R. Ling & P. E. Pedersen (Eds.), *Mobile communications: Renegotiation of the social sphere* (pp. 335–349). London: Springer-Verlag London Limited.
- Ling, Rich, & Baron, N. S. (2007). Text messaging and IM linguistic comparison of American college data. *Journal of Language and Social Psychology*, 26(3), 291–298. doi:10.1177/0261927X06303480
- Ling, Rich, & Baron, N. S. (2013). Mobile phone communication. In S. Herring, D. Stein, & T. Virtanen (Eds.), *Pragmatics of computer-mediated communication* (pp. 191–215). Berlin: De Gruyter Mouton.
- Lipka, L. (2007). Lexical creativity, textuality and problems of metalanguage. In J. Munat (Ed.), (pp. 3–12). Amsterdam: John Benjamins Publishing Company.
- Liwei, G. (2001). Digital age, digital English. *English Today*, 17(3), 17–23. doi:10.1017/S0266078401003030
- Lo, S.-K. (2008). The nonverbal communication functions of emoticons in computer-mediated communication. *CyberPsychology & Behavior*, 11(5), 595–597.

- Morgan, J. (2011, April 8). Why did LOL infiltrate the language? *BBC News Magazine*. Retrieved from <http://www.bbc.co.uk/news/magazine-12893416>
- Murphy, A. P. (2010, September 21). How to decode slang your teen uses online. *ABC Good Morning America*. Retrieved from <http://abcnews.go.com/GMA/Parenting/webspeak-101-parents-decode-teen-internet-slang/story?id=11684997>
- Murphy, A. P., & Allen, J. (2007, January 25). Webspeak: The secret language of teens. *ABC Good Morning America*. Retrieved from <http://abcnews.go.com/GMA/AmericanFamily/story?id=2820582&page=1>
- Nardi, B. A., Whittaker, S., & Bradner, E. (2000). Interaction and outerraction: instant messaging in action. In *Proceedings of the 2000 ACM conference on Computer Supported Cooperative Work* (pp. 79–88). Philadelphia, Pennsylvania, United States: ACM. doi: 10.1145/358916.358975
- North, S. (2007). 'The voices, the voices': Creativity in online conversation. *Applied Linguistics*, 28(4), 538–555.
- Nunes, M. (1997). What space is cyberspace? The Internet and virtuality. In D. Holmes (Ed.), *Virtual politics: Identity and community in cyberspace* (pp. 163–178). London: Sage Publications.
- Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.
- Paolillo, J. C., & Zelenkauskaitė, A. (2013). Real-time chat. In S. Herring, D. Stein, & T. Virtanen (Eds.), *Pragmatics of computer-mediated communication* (pp. 109–133). Berlin: De Gruyter Mouton.
- Pew Research Center. (2005). Internet: The mainstreaming of online life. In *Trends 2005* (pp. 56–69). Washington D.C.: Pew Research Center. Retrieved from <http://pewresearch.org/pubs/206/trends-2005>
- Picasso, P., & Sabartés, J. (1946). *Paintings and drawings of Picasso, with a critical survey*. Paris: Braun & cie.
- PinkNews.co.uk. (2007, July 5). Iraqi death squads use chat rooms to entrap gay men. *PinkNews: Europe's Largest Gay News Service*. Retrieved from <http://www.pinknews.co.uk/news/articles/2005-4861.html/>
- Plag, I. (2003). *Word-formation in English*. Cambridge: Cambridge University Press.

- Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.
- Rainie, L. (2012). *Changes to the way we identify Internet users: Counting Internet users*. Retrieved from Pew Internet & American Life Project website: <http://www.pewinternet.org/Reports/2012/Counting-internet-users/Counting-internet-users.aspx>
- Reid, E. M. (1991). *Electropolis: Communication and community on Internet Relay Chat* (Adapted from an Honor's Thesis, University of Melbourne, Australia). Retrieved from <http://www.irchelp.org/irchelp/misc/electropolis.html>
- Rideout, V. J., Foehr, U. G., & Roberts, D. F. (2010). *Report: Generation M2: Media in the lives of 8- to 18-year-olds* (pp. 1–81). Retrieved The Henry J. Kaiser Family Foundation website: <http://www.kff.org/entmedia/8010.cfm>
- Rosso, M. (2008). User-based identification of Web genres. *Journal of the American Society for Information Science and Technology*, 59(7), 1053–1072.
- Rúa, P. L. (2007). Keeping up with the times: Lexical creativity in electronic communication. In J. Munat (Ed.), *Lexical creativity, texts and contexts* (pp. 137–159). Amsterdam: John Benjamins Publishing Company.
- Sager, J. C. (1990). *A practical course in terminology processing*. Amsterdam ;Philadelphia: J. Benjamins,.
- Samuel, E. (2011, January 11). WoW online game breaks another record. *International Business Times*. Retrieved from <http://www.ibtimes.com/wow-online-game-breaks-another-record-253527>
- Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. London: John Wiley & Sons.
- Shortis, T. (2007). Revoicing txt: Spelling, vernacular orthography and ‘unregimented writing.’ In S. Posteguillo, M. J. Esteve, & M. L. Gea Valor (Eds.), *The texture of the Internet: netlinguistics* (pp. 2–23). Newcastle: Cambridge Scholars Publishing.
- Silva, C. (2010). Chat discourse. In R. Taiwo (Ed.), *Handbook of research on discourse behavior and digital communication: Language structures and social interaction* (pp. 266–280). Hershey, PA: Information Science Reference.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

- Sinclair, J. (2005a). Corpus and text - Basic principles. In A. and H. D. Service & M. Wynne (Eds.), *Developing linguistic corpora: a guide to good practice* (pp. 1–16). Oxford: Oxbow.
- Sinclair, J. (2005b). How to build a corpus. In A. and H. D. Service & M. Wynne (Eds.), *Developing linguistic corpora: a guide to good practice* (pp. 79–83). Oxford: Oxbow.
- Sotillo, S. M. (2010). SMS texting practices and communicative intention. In R. Taiwo (Ed.), *Handbook of research on discourse behavior and digital communication: Language structures and social interaction* (pp. 252–265). Hershey, PA: Information Science Reference.
- Spagnolli, A. (2012). Pragmatics of short message service. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Blackwell Publishing Ltd.
- Sproull, L., & Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communications. *Management Science*, 32(11), 1492–1512.
- Swales, J. M. (1990). The concept of genre. In *Genre analysis: English in academic and research settings* (pp. 33–67). Cambridge: Cambridge University Press.
- Tagg, C. (2009). *A corpus linguistics study of SMS text messaging* (Doctoral dissertation, University of Birmingham, Alabama). Retrieved from <http://etheses.bham.ac.uk/253/>
- Tagliamonte, S. A., & Denis, D. (2008). Linguistic ruin? Lol! Instant messaging and teen language. *American Speech*, 83(1), 3–34. doi:10.1215/00031283-2008-001
- Thurlow, C. (2003). Generation txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online*, 1(1). Retrieved from <http://extra.shu.ac.uk.libproxy.lib.unc.edu/daol/articles/v1/n1/a3/thurlow2002003-paper.html>
- Thurlow, C., & Poff, M. (2013). Text messaging. In S. Herring, D. Stein, & T. Virtanen (Eds.), *Pragmatics of computer-mediated communication* (pp. 135–161). Berlin: De Gruyter Mouton.
- Toms, E. G. (2001). Recognizing digital genre. *Bulletin of the American Society for Information Science and Technology*, 27(2). Retrieved from <http://www.asis.org.libproxy.lib.unc.edu/Bulletin/Dec-01/toms.html>

- Varnhagen, C. K., McFall, G. P., Pugh, N., Routledge, L., Sumida-MacDonald, H., & Kwong, T. E. (2010). Lol: new language and spelling in instant messaging. *Reading and Writing, 23*(6), 719–733. doi:10.1007/s11145-009-9181-y
- Virtanen, S. (2013). Performativity in computer-mediated communication. In S. Herring, D. Stein, & T. Virtanen (Eds.), *Pragmatics of computer-mediated communication* (pp. 269–290). Berlin: De Gruyter Mouton.
- Waldvogel, J. (2007). Greetings and closings in workplace email. *Journal of Computer-Mediated Communication, 12*(2). Retrieved from <http://jcmc.indiana.edu/libproxy.lib.unc.edu/vol12/issue2/waldvogel.html>
- Walther, J. B. (1992). Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication Research, 19*(1), 52–90. doi:10.1177/009365092019001003
- Walther, J. B. (1996). Computer-mediated communication: impersonal, interpersonal, and hyperpersonal interaction. *Communication Research, 23*(1), 3–43. doi:10.1177/009365096023001001
- Walther, J. B., Loh, T., & Granka, L. (2005). Let me count the ways: The interchange of verbal and nonverbal cues in computer-mediated and face-to-face affinity. *Journal of Language and Social Psychology, 24*(1), 36–65. doi:10.1177/0261927X04273036
- Walther, J., & D’Addario, K. P. (2001). The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review, 19*(3), 324–347.
- Waskul, D., & Douglass, M. (1997). Cyberself: The emergence of self in on-line chat. *The Information Society, 13*(4), 375–397.
- Wellman, B. (2004). Connecting communities: On and offline. *Contexts, 3*(4), 22–28. doi:10.1525/ctx.2004.3.4.22
- Werry, C. (1996). Linguistic and interactional features of Internet Relay Chat. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives* (pp. 47–63). Amsterdam: J. Benjamins.
- Wildemuth, B. M. (2009). Frequencies, cross-tabulation, and the chi-square statistic. In B. M. Wildemuth (Ed.), *Applications of social research methods to questions in information and library science* (pp. 348–360). Westport, Connecticut: Libraries Unlimited.

- Wilkins, H. (1991). Computer talk: Long-distance conversations by computer. *Written Communication*, 8(1), 56–78. doi:10.1177/0741088391008001004
- Wutiolarn, N., & Attaprechakul, D. (2012). A study of nonstandard orthography and vowel omission in an international online game: AuditionSEA. In *English language research: ASEAN synergy of pedagogical and professional perspectives The First LITU International Graduate Conference* (pp. 89–97). Bangkok, Thailand: Thammasat University. Retrieved from http://litu.tu.ac.th.libproxy.lib.unc.edu/2012/images/litu/pdf/Total_Proceeding.pdf#page=89
- Yates, S. J. (1996). Oral and written linguistic aspects of computer conferencing: A corpus based study. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives* (pp. 29–46). Amsterdam: John Benjamins Publishing Co.
- Yee, N. (2006). The labor of fun: How video games blur the boundaries of work and play. *Games and Culture*, 1(1), 68–71. doi:10.1177/1555412005281819
- Yongyan, L. (2000). Surfing e-mails. *English Today*, 16(4), 30–55.
- Zawada, B. E. (2005). *Linguistic creativity and mental representation with reference to intercategory polysemy* (Doctoral dissertation, University of South Africa). Retrieved from <http://uir.unisa.ac.za.libproxy.lib.unc.edu/handle/10500/1965>