

Who's Blogging Now?

Linguistic Features and Authorship Analysis in Sports Blogs

by

Taylor Cox

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2017 by the
Graduate Supervisory Committee:

Elly Van Gelderen, Chair
Carrie Gillon
Elisabeth Gee

ARIZONA STATE UNIVERSITY

December 2017

ABSTRACT

The field of authorship determination, previously largely falling under the umbrella of literary analysis but recently becoming a large subfield of forensic linguistics, has grown substantially over the last two decades. As its body of research and its record of successful forensic application continue to grow, this growth is paralleled by the demand for its application. However, methods which have undergone rigorous testing to show their reliability and replicability, allowing them to meet the strict Daubert criteria put forth by the US court system, have not truly been established.

In this study, I set out to investigate how a list of parameters, many commonly used in the methodologies of previous researchers, would perform when used to test documents of bloggers from a sports blog, *Winging It in Motown*. Three prolific bloggers were chosen from the site, and a corpus of posts was created for each blogger which was then examined for each of the chosen parameters. One test document for each of the three bloggers which was not included in that blogger's corpus was then chosen from the blog page, and these documents were examined for each of the parameters via the same methodologies as were used to examine the corpora. Once data for the corpora and all three test documents was obtained, the results were compared for similarity, and an author determination was made for each test document along each parameter.

The findings indicated that overall the parameters were quite unsuccessful in determining authorship for these test documents based on the author corpora developed for the study. Only two parameters successfully identified the authors of the test documents at a rate higher than chance, and the possibility exists that other factors may be driving these successful identifications, demanding further research to confirm their validity as parameters for the purpose of authorship work.

ACKNOWLEDGMENTS

I am grateful to my chair, Elly Van Gelderen, who helped me keep my head on straight in order to shape the course of my graduate studies, my research project, and this dissertation itself. I will never forget her displays of patience and encouragement. I do not know how I would have made it through without her. I am also thankful to my committee members, Betty Gee, who stepped in unexpectedly to fill the hole of a lost committee member and has been an inspirational source of encouragement and direction ever since, and Carrie Gillon, whose knowledge and expertise I have had the privilege to have access to since my undergraduate experience and on whom I have always been able to rely, even in difficult times.

I owe a special thank you to Dr. Mark Davies of Brigham Young University for aiding me in my attempts to acquire parts-of-speech counts from his Corpus of Contemporary American English for comparison in the register analysis portion of this study. He kindly and swiftly stepped in to provide me this information, allowing me to more accurately analyze my own data.

I must also thank my long list of family and friends, who have been standing by, waiting a long while for me to finish this degree so that I might be able to come out and play a little more often. I cannot possibly name you all here, but if you are wondering if this means you, it definitely does. My mother has supported me throughout this process in ways I simply cannot innumerate here. I owe a special thank-you to my dear friend Dana Wakiji for reviewing many documents, providing unending advice and supportive words, and encouraging me non-stop throughout this process. Thank you for our safe space. My beer buddies and hockey pals have supported and encouraged me through this process more than they may ever know- we are about to celebrate epically, you guys.

TABLE OF CONTENTS

	Page
LIST OF TABLES	ix
LIST OF IMAGES	xiii
CHAPTER	
1 INTRODUCTION	1
1. Introduction.....	1
2. Purpose of the study	2
3. Overview of chapters	3
2 WHAT'S IN A BLOG?.....	6
1. Background of blogs	6
2. History of blogs.....	6
3. What is a blog?	9
4. Who writes and reads blogs?	11
4.1 Blogger demographics.....	12
4.2 Reader demographics	13
5. Types of blogs.....	14
6. What about Winging It in Motown?	17
3 SITUATIONAL CHARACTERISTICS OF WINGING IT IN MOTOWN	19
1. Introduction.....	19

CHAPTER	Page
2. Background Information on Blogs.....	20
2.1 Early History of Blogs.....	20
2.2 Frequently-Cited Characteristics of a Blog.....	21
3. Biber and Conrad’s Register Analysis as a Framework.....	21
4. Form	22
4.1 Examining a blog’s form.....	22
4.2 Post frequency	23
4.3 Home Page	24
4.4 Beyond the Home Page.....	27
5. Situational Characteristics.....	32
5.1 Participants	33
5.2 Relations among participants	37
5.3 Channel	50
5.4 Production Circumstances	52
5.5 Setting	53
5.6 Communicative Purpose	55
5.7 Topic.....	68
4 LINGUISTIC FEATURES OF WINGING IT IN MOTOWN	69
1. Introduction.....	69

CHAPTER	Page
2. Creating a corpus.....	69
3. Word Count	71
4. Sentence count and length	72
5. Parts of speech counts	72
5.1 Nouns and Verbs	73
5.2 Other Parts of Speech	79
6. Lexical Information.....	91
7. Biber’s MAT analysis.....	93
7.1 Dimension 1.....	93
7.2 Dimension 2.....	94
7.3 Dimension 3.....	94
7.4 Dimension 4.....	95
7.5 Dimension 5.....	96
7.6 Dimension 6.....	97
5 WHO WROTE IT? THE AUTHORSHIP STUDY	98
1. Background.....	98
1.1 Early authorship work	98
1.2 Modern Research	101
2. Experiment design.....	105

CHAPTER	Page
2.1 Methodology	105
3. Results	112
3.1 Average sentence and word length	113
3.2 Type-Token Ratio	113
3.3 Readability- Readability Test Tool	113
3.4 Readability- Online Utility	113
3.5 Parts of Speech.....	114
3.6 Modals.....	115
3.7 Function word profile.....	116
3.8 Word length profile- length frequency counts and percentages.....	117
3.9 Word length profile- frequency ranking	119
3.10 Punctuation mark profile- raw count and normalized per 100 words.....	120
3.11 Verb forms- raw counts and counts per 100 verbs/words	121
3.12 Syllable counts, percentage relative to total word count.....	122
3.13 Grammatical errors, raw count and normalized to per 100 words	123
3.14 Spelling errors	123
3.15 Syntactically-classified punctuation- raw counts.....	124
3.16 Syntactically-classified punctuation- normalized.....	125
3.17 Grapheme (unigram) profiles.....	127

CHAPTER	Page
3.18 Character n-grams, direct correlation in position	128
3.19 Character bigrams and trigrams, top 10, 20, and 50 in common	129
4. Identifications per parameter	130
4.1 Average lengths.....	131
4.2 Type-token ratio.....	131
4.3 Lexical density	131
4.4 Readability- Readability Test Tool.....	131
4.5 Readability- Online Utility	132
4.6 Parts of Speech.....	132
4.7 Modals.....	133
4.8 Function word profile comparison, correlation of function word ranking ..	133
4.9 Word length count total comparison, frequency.....	134
4.10 Word length profile total comparison, direct correlation of ranking	134
4.11 Punctuation profile comparison.....	134
4.12 Verb forms.....	135
4.13 Syllable count comparison	135
4.14 Grammatical errors	135
4.15 Spelling errors	136
4.16 Syntactically classified punctuation	136

CHAPTER	Page
4.17 Grapheme profile comparison	137
4.18 Character n-gram comparison, direct correlation of position	137
4.19 Character bigram and trigram comparisons, top in common.....	138
5. Overall identification success per parameter	138
6 WHAT DOES IT ALL MEAN, AND WHAT IS THE NEXT STEP?	140
1. Summary and discussion of results	140
2. Challenges and limitations of the study	143
3. Future directions	145
REFERENCES	147
APPENDIX	
I TEST DOCUMENTS	151

LIST OF TABLES

Table	Page
4.1 Nouns	73
4.2 Verbs	74
4.3 Other Parts of Speech	80
4.4 Adjectives.....	81
4.5 Adverbs.....	81
4.6 Pronouns	81
4.7 Wh- Words	82
4.8 Top 10 Words.....	92
4.9 Top 10 Non-Function Words	93
5.1 Average Sentence and Word Length.....	113
5.2 Type-Token Ratio	113
5.3 Readability- Readability Test Tool	113
5.4 Readability- Online Utility	114
5.5 Parts of Speech	115
5.6 Modals	116
5.7 Function Word Profiles.....	116

Table	Page
5.8 Word Length Profiles- Frequency Counts.....	118
5.9 Word Length Profiles- Ranking.....	120
5.10 Punctuation Mark Profiles	121
5.11 Verb Forms	122
5.12 Syllable Counts.....	123
5.13 Grammatical Errors	123
5.14 Spelling Errors.....	124
5.15 Syntactically Classified Punctuation- Raw	125
5.16 Syntactically Classified Punctuation- Normalized	127
5.17 Grapheme (Unigram) Profile.....	128
5.19 Bigrams- Direct Correlation in Position.....	128
5.20 Trigrams- Direct Correlation in Position.....	128
5.21 Top 10 Bigrams	129
5.22 Top 20 Bigrams	129
5.23 Top 50 Bigrams	129
5.24 Top 10 Trigrams	129
5.25 Top 20 Trigrams	130

Table	Page
5.26 Top 50 Trigrams	130
5.27 Average Lengths Comparison	131
5.28 Type-Token Ratio Comparison	131
5.29 Lexical Density Comparison.....	131
5.30 Readability Comparison- Readability Test Tool	132
5.31 Readability Comparison- Online Utility	132
5.32 Parts of Speech Comparison	133
5.33 Modal Comparison	133
5.34 Function Word Profile Comparison.....	133
5.35 Word Length Count Total Comparison- Frequency	134
5.36 Word Length Comparison- Direct Correlation of Ranking	134
5.37 Punctuation Profile Comparison.....	134
5.38 Verb Form Comparison	135
5.39 Syllable Count Comparison.....	135
5.40 Grammatical Errors Comparison	136
5.41 Syntactically-Classified Punctuation Comparison	137
5.42 Character N-Gram Comparison	138

Table	Page
5.43 Top Bigram Comparisons	138
5.44 Top Trigram Comparisons	138
5.45 Accuracy	139

LIST OF IMAGES

Image	Page
3.1: Home Page 1	24
3.2: Home Page 2	25
3.3: Home Page 3	26
3.4: Home Page 4	27
3.5: Sections Page	28
3.6: Fanposts Page	29
3.7: Fanshots Page	29
3.8: A Typical Blog Post 1	30
3.9: A Typical Blog Post 2	31
3.10: Typical Comments Section on a Blog Post	32
3.11: The Masthead.....	33
3.12: Blogger JJ Interacting in the Comments.....	37
3.13: 840 Comments on a Gameday Thread.....	38
3.14: 474 Comments on a Post About a Trade	38
3.15: A Typical Gamethread	39
3.16: Start of a Gamethread.....	40
3.17: Gamethread Reaction to Live Event.....	40

Image	Page
3.18: Gamethread Reaction to Live Event.....	41
3.19: Gamethread Reaction to Live Event.....	41
3.20: Fanshot 1	42
3.21: Fanshot 2	42
3.22: Fanpost 1	43
3.23: Fanpost 2	43
3.24: Blogger JJ Threatening Disciplinary Action in Comments	44
3.25: Personal Interaction in the Gamethread Comments.....	45
3.26: Presumption of Shared Knowledge 1	47
3.27: Presumption of Shared Knowledge 2	48
3.28: Presumption of Shared Knowledge 3	49
3.29: Chart in a Blog Post	50
3.30: Embedded Videos/GIFs in a Blog Post	51
3.31: Link to External Content- Informative.....	55
3.32: Link to External Content- Editorialized	56
3.33: Informative- Getting to Know 1	57
3.34: Informative- Getting to Know 2	57

Image	Page
3.35: Analytics/Statistics Post.....	58
3.36: Pregame Post.....	58
3.37: Post Expressing Blogger Opinion.....	59
3.38: Post Expressing Blogger Opinion.....	59
3.39: Post Expressing Blogger Opinion.....	59
3.40: Post Critical of Mainstream Beat Writer 1	60
3.41: Post Critical of Mainstream Beat Writer 2	61
3.42: Post Critical of Mainstream Beat Writer 3	61
3.43: Petrella’s Criticism of the Detroit Mainstream Media 1.....	62
3.44: Petrella’s Criticism of the Detroit Mainstream Media 2.....	63
3.45: Petrella’s Criticism of the Detroit Mainstream Media 2.....	63
3.46: Petrella’s Criticism of the Detroit Mainstream Media 2.....	63
3.47: Petrella’s Criticism of the Detroit Mainstream Media 3.....	64
3.48: KyleWiiM’s Opinion Piece 1	65
3.49: KyleWiiM’s Opinion Piece 2.....	65
3.50: Prediction Post 1.....	66
3.51: Prediction Post 2.....	66

Image	Page
3.52: Prediction Post 3.....	67
3.53: Prediction Post 4.....	67

Chapter 1

Introduction

1. Introduction

A significant amount of communication today takes place online, in a variety of formats. The purposes of these communications are countless and they can cover virtually any topic. With so much daily communication occurring in this manner, circumstances sometimes arise wherein such communications become the subject of evidentiary interest to law enforcement and related agencies, both for investigative and for court purposes. These circumstances sometimes lead law enforcement personnel to turn to linguists for expert knowledge, including regarding assistance with the analysis of texts. This is especially true in cases of unknown or disputed authorship.

Law enforcement officials sometimes consult linguists working in the field of forensic linguistics, the intersection of linguistic study and the law, to help determine or verify the authorship of disputed texts. While authorship analysis, including verification and identification, was among the first subfields to emerge under the umbrella of forensic linguistics, reliable and replicable methodologies that meet the standards set forth by many courts, including the Daubert standard adhered to in US courts, have not truly been established and remain a challenge. The task set before forensic linguists working with authorship approaches remains to discover methods that have high reliability rates and low error rates, that can be uniformly applied to numerous types of text data (e.g., from numerous genres, by authors falling into different demographics, and so on) with relatively the same rate of reliability, and that specifically can handle not only long texts providing a strong amount of data to

work with but also more forensically-realistic shorter texts, which have proven to be an especially significant challenge in the field. Many previously-studied methods lack the robust base of replicable research results supporting their success required to meet the stringent standards set forth by the courts.

2. Purpose of the study

This study is aimed at strengthening the body of work available in the field of forensic authorship identification by providing data on an understudied register of text, in this case a single sports blog with multiple authors. Further, the study adds data on the performance of multiple parameters of comparison such that their reliability may be further examined and the bounds of that reliability tested. Specifically, my approach for this study was to utilize a number of previously-tested techniques, some of which are commonly discussed as possible identifying parameters in the areas of linguistic and stylistic authorship identification, to attempt to identify the authors of texts in the form of blog posts culled from the ice hockey blog *Winging It in Motown*, also known as *WiiM*. This also demanded a register analysis of the blog from which the posts were drawn, such that its situational characteristics and linguistic features could be determined. My ultimate goal was to determine which parameters are most effective in achieving these determinations, thereby adding to the data on each parameter's reliability across registers.

Research questions:

- A. What are the linguistic characteristics of sports blogs?
- B. How do these characteristics relate to those found in other registers by Biber and Conrad?

- C. Which technique for authorship identification is most reliable on text derived from a sports blog?

3. Overview of chapters

Chapter 1 is an introduction to this dissertation which lays out the purpose of the primary study as well as an overview of each of the six chapters.

Chapter 2 provides an overview of the background of blogs as a text type. This overview includes a discussion of the history of blogs, charting the evolutionary course of the register as it developed over a period of two and a half decades beginning in the early 1990s and leading up to present day. A discussion is presented of what defines a blog, including common characteristics and how the definition has evolved as the register has grown. The question of what constitutes a blog naturally leads to the question of who writes blogs and who reads them, and an overview of available surveys of the demographics of both bloggers and blog-readers is presented to help answer this question. A discussion of the various types of blogs commonly found on the internet today follows. Finally, *Winging It in Motown*, the blog at the center of this dissertation, is briefly situated within the blog overview that has been presented in this chapter.

The discussion of how *WiiM* fits in among blogs leads directly into chapter 3. In order to examine the language of participants on the blog, the blog must be understood as a register. Chapter 3 presents the first part of Douglas Biber's two-part register analysis, as conducted on *WiiM*. The first part, discussed in this chapter, is an analysis of the form of *WiiM*, following the framework presented by Biber and Conrad (2009). A brief introduction to the register framework is given, and then the framework is covered step by step with *WiiM* as the subject. The primary focus of

this first part of the analysis is on the form and situational characteristics of the text. I first examine the form, discussing post frequency as well as the design of the blog's home page, which is extensive. I cover the design of other primary pages on the blog and then provide an overview of the design of a typical blog post on the page. I then move on to covering the situational characteristics of the blog, as outlined by Biber and Conrad (2009). These characteristics include participants, relations among participants, channel, production circumstances, setting, communicative purpose, and topic.

Chapter 4 covers the second half of the register analysis process. In this chapter, I discuss the process of creating a corpus of blog posts from WiiM, and then work through the analysis of the linguistic features of the blog according to Biber and Conrad's (2009) methodology. The process begins with a word count, both for the entire corpus and to determine an average words-per-post count for comparison to previous averages found. The next step is a count of sentences in the corpus and a determination of their average length in words. Once these counts have been taken, an analysis of the parts of speech composition of the corpus is carried out. This includes counts of different categories of nouns, verbs, adjectives, adverbs, pronouns, and wh-words, as well as prepositions and subordinating conjunctions, modal auxiliaries, determiners, interjections, coordinating conjunctions, existential *there*, foreign words, and cardinal numbers. The counts are presented, discussed, and analyzed within the context of the blog as a register. Counts found by Biber et al. (1999) and in the Corpus of Contemporary American English (COCA), as provided by Mark Davies via private correspondence (Davies, personal communication, April 29, 2017), are provided for reference regarding expectations of average counts across registers and in English in general. A discussion of lexical information follows,

including the top 10 words overall as well as the top 10 lexical words. The chapter concludes with the corpus data being run through Biber's MAT analysis program. The results for each dimension are presented and discussed.

Chapter 5 shifts the focus to the main study of the dissertation, centered on the forensic linguistics subfield of authorship identification. This chapter includes a short overview of the history of authorship identification focused specifically on forensic authorship work, with early history being presented chronologically and modern studies broken down by approaches and major contributors. Next, the methodology of the study is discussed. I discuss the process of building author corpora, including choosing which bloggers to include, as well as the processes of choosing parameters to examine, obtaining baseline measurements of these parameters per author via the author corpora, and then choosing test documents and examining them via the same parameters. The processes required for examining each parameter are covered in detail. The results for each parameter are then presented for all three author corpora and all three test documents side by side, for ease of reader comparison. Finally, author identifications for each test document based on each parameter are presented, and then the accuracy of each parameter's determination and a statement of whether the accuracy is better than chance are given.

The final chapter, chapter 6, begins with a summary of this dissertation followed by a discussion of the results and their implications for the study and for the field. Following this, I cover the limitations of the study. The discussion of limitations leads into a discussion of future directions for research, including how the research can be expanded using this data set.

Chapter 2

What's in a Blog?

1. Background of blogs

Before beginning either register analysis or authorship work using blogs, the answer to one crucial question must be established: what is a blog? This question will be answered via a discussion of the background of blogs as discussed in previously-existing research, including a history of blogs as a text type as well as an explanation of different types of blogs and their purposes. This discussion will offer an overview of blogs as a whole, which will both help define WiiM as a blog and situate it as a subregister against other types of blogs.

2. History of blogs

The advent of the early iteration of the internet in the 1960s changed the course of technological history (Crystal, 2009). Though few had access at that point or were even aware of the existence of such technology, that advancement led to an eventual increasingly rapid evolution of technology, with modern technology not even conceivable just a few decades ago available at the fingertips of almost anyone today. In 1983 and 1984, *USENET* and *Listserv*, respectively, were developed to facilitate online communication, and in 1992, Tim Berners-Lee developed the internet's first website (Carvin, 2007). Though these early precursors lay the foundation for the ability to develop pages on the internet for hosting interactive communication, including blogs, the first major evolutionary occurrence in the development of the blog world was Pyra's 1999 launch of a platform designed specifically to facilitate the building of personal blog pages by users with no coding skills (Stone, 2004). According to Stone (2004), the platform in fact began as an

attempt to build a project management platform, with no original intention toward blogging or any similar activity. However, Pyra would find themselves launching what ultimately ended up being one of the heaviest-used personal blogging platforms on the internet, eventually termed *Blogger*. Stone (2004) and other blog scholars such as Rebecca Blood (2000), one of the earliest scholars to begin researching blogs as a communicative medium, often credit the conception of *Blogger* as the event that kicked off the blogging revolution by bringing the ability to publish a personal blog to virtually anyone with internet access, rather than just coding experts. During the same year that Pyra launched *Blogger*, the lesser-known *Pitas* and *Groksoup* were also released as free web-based build-your-own-blog tools (Blood, 2000). The launch of these three platforms opened the floodgates for the publishing of personal blogs.

While the development of user-friendly blogging platforms was occurring, individual web users with coding skills were beginning to develop and publish their own blog pages, leading to the creation of terms which are considered standard now. According to Blood (2000), Jorn Barger coined the term *weblogs* in 1997. As readers began to visit the blog pages published by these coders, more began to participate and to link to each other's pages. One member of this community, Peter Merholz, declared his intention to pronounce *weblog* as *wee-blog* instead of *web-log*, leading to the familiar modern clipping *blog* (Blood, 2000).

In 2003, Google acquired *Blogger*, giving the site major financial and recognition backing and access to a large group of top developers, cementing its position as a central feature of web use. Meanwhile, the platform that would become the other major personal blog building site alongside *Blogger*, *Livejournal*, was developed and launched around the turn of the century, and saw a rapid increase in users of its own, numbering in the millions within several years of launch (Stone,

2004). Several other sites which would become popular build-a-blog sites launched between 2000 and 2003 as well, most notably *Xanga* and *Wordpress*, further facilitating the growth of the blogging revolution (Carvin, 2007). The development of these sites also led to a transition in the use of blogs, from their origins as largely just simple collections of links to web pages all displayed in one spot to the diary-style postings seen on sites like *Livejournal* and *Xanga* and eventually to more topic-driven informational blogs such as the blog used in this study (Salen, 2007, pp. 32). These types will be further described and discussed later in this chapter.

In the years that followed, the use of blogs continued to grow. The range of topics covered by blogs grew as well, and some bloggers even began to be granted official credentials for access in their blog's topic field (Carvin, 2007). By 2007, the number of blog pages published on the internet eclipsed the 100 million mark (Carvin, 2007). The web platform *Twitter* was launched in 2006, inspiring the use of the term *microblog*, and by 2012 it was ranked as one of the top 10 most visited sites on the web with upwards of half a billion active users (Walker Rettberg, 2014). A second popular microblogging platform, *Tumblr*, launched one year later, in 2007, securing the popularity of the microblog alongside more traditional blog types. Walker Rettberg (2014) argued that even some social media sites such as *Facebook*, not typically considered a blog site, are "at root a form of blogging," and that the idea behind the entire concept of social media usage boils down to the same as that of blogs: "let everybody share their thoughts and discoveries online" (pp. 14).

As Walker Rettberg (2014) noted, determining the number of blogs on the internet is a monumental task, with no central counting agency and with blogs spread to the far corners of the internet and occurring in a variety of formats. Determining what qualifies as a blog, including whether microblog accounts should

be included, and then developing the means to count them are logistically daunting tasks, and this has led to a lack of reliable statistics. Counts that have been attempted, outdated by several years now, suggested the number to be into the hundreds of millions and trending upward (Nielsen, 2012; Statista, 2017b). Statista listed the number of microblogs on *Tumblr* at over 345 million in April 2017 and the number of active *Twitter* users at 328 million in the first quarter of 2017 (Statista, 2017a, 2017c).

3. What is a blog?

In order to study blogs in any sense, be it an in-depth linguistic analysis or a simple count, what constitutes a blog must first be established. Blogs come in a wide variety of shapes and sizes, and one can randomly select two blogs for comparison and find that they bear little resemblance to each other. However, there are some characteristics common across blogs. These characteristics are not necessarily required, but are found in the majority of blogs and, taken in clusters, generally reliably indicate whether a site is considered a blog by its creator and its users.

Rebecca Blood (2002a), one of the first to attempt to characterize blogs, defined them as “a frequently updated webpage with dated entries, new ones placed on top,” a format which she postulated was chosen “as a matter of convenience, so that visitors could instantly see their latest update, and whether it had been made a week, a day, or an hour ago” (pp. ix). Susan Herring et al. (2005) similarly defined blogs as “frequently modified web pages in which dated entries are listed in reverse chronological sequence” (pp. 142) “‘Links with commentary, updated frequently’ was the formula” according to Blood (2002a, pp. ix). Indeed, the majority of blogs have dated entries listed in reverse chronological order, and whether they are updated

regularly or not, users following them generally expect them to be. The characteristic of a list of dated entries in reverse chronological order is among the most commonly-cited qualities characterizing a webpage as a blog (Herring, Scheidt, Wright, & Bonus, 2005; Bar-Ilan, 2005; Schmidt, 2007; Brala, 2008; de Moor & Efimova, 2004).

The earliest blogs were primarily lists of web links the blog author wished to share, sometimes with commentary and sometimes with little to none, and eventually they came to include links to other blogs as well (Myers, 2010). As Crystal (2009) noted, "links are very important" as a defining feature of blogs, and "some blogs consist of little more than a long list of hyperlinks" (pp. 240). The inclusion of numerous hyperlinks to other web content is thus a second heavily-cited characteristic of blogs. As Blood (2000) stated, "the original weblogs were link-driven sites" and consisted of "a mixture in unique proportions of links, commentary, and personal thoughts and essays." As the community utilizing blog pages began to grow and spread beyond the realm of the technologically savvy, the purpose of blogs also expanded. Not only were these collections of links published to share pages the blog author liked, but also to filter web content about specific topics for readers.

According to Blood (2004), early blogs were "rudimentary in design and content" (pp. 54). As the popularity of blog use, both creating and consuming, continued to grow, however, the complexity of features and characteristics of blogs grew in kind. Per Schmidt (2007), authors may utilize a variety of content formats to create their blogs, including text, images, and sound files. Nowson, Oberlander, and Gill (2005) stated that blogs "contain[s] news and views on a variety of topics" and "are already seen as a powerful news-gathering medium," suggesting substantial growth in both complexity and purpose (p. 1666).

4. Who writes and reads blogs?

Blogging demographics are somewhat surprisingly hard to come by. Most available studies are fairly old, carried out primarily between 2008 and 2012. Examining blog demographics is a rather mighty task, particularly when examining demographics across blogs for a more representative sample of bloggers and blog readers in general, rather than examining the authors and readers of a single blog. Such a study is necessarily driven by self-report, presenting another potential methodological issue, as one must rely on responders to report demographic information honestly and accurately and to understand questions and prompts appropriately. Furthermore, with millions of blogs scattered across the internet and the issue of whether to consider microblogs as well, covering every single blog would be a logistically impossible task. Few researchers have even undertaken the task of attempting to examine the demographics of the most heavily visited blog sites on the internet in the hope that a representative sample could be derived. Even that task is monumental, requiring the agreement of the owners of the blogs and then a willingness on the part of bloggers and readers to take part in a survey. A further issue is that some bloggers work diligently to deliberately keep their identities hidden, making their demographics difficult to access and leading them to likely be uncooperative in any efforts to collect such information (Dardick, La Roche, & Flanigan, 2007). Because of these obstacles, little information was available, but data from reliable sources with studies that appear to be as methodologically sound as possible was gathered and is presented here to attempt to offer a basic overview of blogger and blog reader demographics.

4.1 Blogger demographics

The demographics of both blog readership and blog authorship have been examined, but the results have varied quite widely. Some examinations have found that the majority of authors are male (Technorati, 2010) while others have found the majority to be female (Nielsen, 2012). Technorati and Nielsen both found that bloggers were a highly-educated group, with upwards of seven out of 10 having attended college and more than four out of 10 holding graduate degrees (Technorati, 2010; Nielsen, 2012). Both Nielsen and Technorati reported the majority of bloggers as being young adults or middle-aged, with Nielsen reporting that half of bloggers are aged 18-34 (2012) and Technorati reporting 65% of bloggers are between 18 and 44 (2010). Nielsen found that about one in three bloggers is a mom and that around 52% are parents with minor children in their home (2012), while Technorati's report showed that about 48% of bloggers reported being parents (2010). Geographically speaking, Technorati also found US bloggers to be relatively evenly-distributed across the country, and that a large portion of the world's bloggers are in North America (though it is worth noting that their study was a self-report survey which was presented in English) (2010). They also found that a large percentage of bloggers were either employed full-time or self-employed, that the vast majority did not earn their full income from blogging, and that most had been blogging for at least two years with around one fifth to one quarter having been blogging for six years or more (2010).

In 2013, Ignitespot posted an infographic of blogging statistics which included some demographic information on bloggers (Hood, 2013). They found the majority of bloggers to be female, as Nielsen did (2012). They also found that 53.3% of bloggers are between the ages of 21 and 35. They found Blogger to be the heaviest-used

blogging site, with 46 million unique visitors monthly. They found that 6.7 million people blog on blogging sites and 12 million blog on social networking sites such as Twitter and Facebook, often also considered microblogs. They also described five types of blogger: the part-time professional, who supplements her income with revenue from blogging activity; the hobbyist, who earns no income from blogging activity, blogs for personal enjoyment, and often posts personal opinions or experiences; the full-time professional, whose primary job and source of income is blogging; the corporate, who blogs for the company or business that employs them; and the entrepreneur, who blogs for their own business. They further found that around 14% of bloggers earn income from their blogging activity.

4.2 Reader demographics

If the demographics of the vast number of bloggers across the internet are difficult to obtain, the demographics of the readers of their blogs are even more so. However, it is valuable to understand who reads blogs as part of what defines the blog as a text type. Ignitespot (Hood, 2013) determined that 77% of internet users read blogs, a significant portion. A 2013 Pingdom study examining a collection of “the world’s top blogs” provided some demographic information on blog readers. They examined 80 blogs of various styles covering various topics, though none were of the type seen on diary-style build-your-own-blog sites like LiveJournal. Their findings showed distribution of demographics to vary drastically across the blogs, with, for example, age demographics ranging from around 40% 18-24-year-olds and no readers over 65 on one blog to less than 5% 18-24-year-olds and close to 25% 65+ on another. They found a median age across blogs of 38 and an average age of 40.7, numbers they found surprisingly high. Gender distribution also varied, with 63/37 split favoring female readers on one end of the spectrum and a 70/30 split

favoring male readers on the other. The study showed an average split of 55% male and 45% female readership. They further found that 59 of the 80 blogs showed a male-dominant readership.

Blogads carried out a survey in 2004 of over 17,000 blog readers. They also found the largest portion of the demographic to be between the ages of 31-40, at 29.4% of responders (Copeland, 2004). About 27% were between the ages of 19-30, and about 37% between 41-60. These results also suggest that a large portion of blog readers are middle-age, again surprising the conductors of the study, who anticipated a younger audience. The respondents to this survey also suggested a staggering male leaning, with 79.1% of respondents stating they were male and only 20.9% stating they were female. This result skewed much higher to the male side of the demographic than Pingdom's later survey. Blogads' survey respondents also skewed democratic in political leaning at over 40% of respondents while only 22.6% claimed to be republican, and the vast majority of the respondents, at 91.4%, were located in the US, with the most represented state being California. The majority of their respondents also appeared to be middle class, with the most common salary range being \$60-90,000, at almost 22% of respondents. Salary ranges on either side of this, \$45-60,000 and \$90-120,000, came in third and second place, respectively.

5. Types of blogs

Blogs can come in different formats and via different media, and exist for different purposes or cover different topics. These all suggest ways in which blogs can be categorized. Many blogs are primarily presented in the format of text, written by the author or perhaps as excerpts or quotations from other written sources. However, blogs can also come in the format of video clips, often referred to as

videoblogs or vlogs, which may contain little or no text as the content of the blog is spoken in the presented video clips (Crystal, 2009). Collections of linked video can also qualify as vlogs. Some blogs may be in the form of audio clips or music, referred to as audioblogs, and some may even simply be collections of photographs, referred to as photoblogs (Crystal, 2009). Blogs can occur as any of these formats, and many include a combination of multiple types of media, making them multimodal in nature. GIFs, compressed files showcasing moving images, much like short video clips, but lacking sound, are often utilized in text-based blogs now as well, adding a new dimension to the multimodality of blogs (merriamwebster.com, 2017).

In addition to the format and media type of blogs, they can be characterized by their purpose and topic as well. As discussed in section three, the earliest blogs were primarily collections of links to other webpages the author wished to share (Myers 2010). The collections may have simply been to showcase pages the author liked and enjoyed, or they may have been collections of topic-driven sites that shared a common subject, and the inclusion of blogger commentary varied from virtually none to a significant discussion of a link. However, regardless of the amount of blogger input, the primary directive of the posts was to share collections of links (Herring, Scheidt, Wright, & Bonus 2005). Blogs centered on links also meet the definition of Blood's (2002b) blog type termed *filter blogs*, which she characterizes as blogs that revolve primarily around links to external web content with the amount of commentary ranging widely.

Once technology companies began developing platforms that removed the need for significant coding in order to publish a blog page, opening up the ability to publish a blog to virtually anyone, the purpose of blog pages began to shift. These platforms, such as *Blogger*, *WordPress*, *Xanga*, and *Livejournal*, led to the concept of

different types of blogs, as with more individuals with access to the ability to publish blogs came a wider variation in their use. *Xanga* and *Livejournal*, in particular, consisted of many blogs of the *diary* type, also referred to as the *personal journal* type, with bloggers using the pages less as a way to curate web content and more as a platform for expressing personal thoughts and feelings as well as personal experiences on a wide variety of subjects, some as mundane as what the blogger prepared for breakfast (Herring, Scheidt, Wright, & Bonus 2005; Garden, 2012). This type of blog is generally referred to as a *diary* blog because of its structural and contextual similarity to the genre of the diary entry. Social media posts falling under the umbrella of microblogs often reflect the characteristics of this type of blog as well. Such blogs are generally not focused on any one topic (Walker Rettberg 2014).

As the complexity of coding schemes available to the average internet user increased, so, too, did the uses and purposes of blogs, as well as their designs and characteristics. *Corporate* blogs have grown in popularity in recent years. Debbie Weil (2006) defined corporate blogs as “the use of blogs to further organizational goals” (pp. 1). These blogs are set up by companies to discuss issues and topics relating specifically to the company itself and its directives, and many companies are now engaged in this practice. As Weil (2006) extolled to her readers, “a blog is a marketing communications channel” (pp. 2) which can help an organization or company meet a variety of its goals and which enables conversations to take place between and among the organization or company, its employees, its customers or consumers, and others in the industry. Generally, either the owner of the company authors the posts or an employee is charged with this task as part of their job duties, though some companies seek outside help in the creation and management of their blog content (Weil, 2006).

Some blogs are topic-centered and driven primarily by a desire to keep readers informed regarding the topic and to allow bloggers to express their thoughts and opinions on the topic in general as well as new developments as they occur and are shared by the blogger. These blogs often follow the style of *filter* blogs as defined by Blood (2002b), with links to news and other web content related to the blog topic frequently a central part of blog posts. Blood specifically noted that some filter blogs “focus on a particular subject” with the goal being “to provide their readers with a continuous source for all the available news about a given topic” (pp. 8). However, while Blood (2002b) discusses filter blogs as being specifically centered on links to external web content, some topic-driven blogs may vary their reliance on links widely. Bloggers may vary their reliance on linked content while creating topic-specific posts by posting both blog posts with links at the center and blog posts focused more on personal stance and thought expression or original analysis, or even sometimes by obtaining news or information regarding the topic first-hand. As the blogosphere expands, blogs can be run as an original source, such as the corporate blogs discussed above, which can generally be considered topic-focused with the topic being the company itself. Furthermore, some bloggers have managed to gain notoriety, credibility, and respectability which has enabled them to gain access to first-hand information on topics as well.

6. What about Winging It in Motown?

The discussion regarding what constitutes a blog and how blogs are categorized leads to the question of how the blog examined in this study, *Winging It in Motown*, fits in to this schema. A deeper analysis of *WiiM*, which will illustrate in more detail how *WiiM* functions as an example of the register of the blog, will be given in chapter 3. However, here a brief overview can be given. In keeping with the

basic definition of a blog, WiiM is a web page that is updated regularly, sometimes several times per day, with dated entries listed in reverse chronological order. WiiM posts often, though not always, include linked content, and this content varies in format, including text, photos, audio, and video. WiiM is a blog that is heavily focused on a specific, narrowly-defined topic, one specific NHL team, with rare deviations to other closely related topics such as other teams in the league. The blog is closest to Blood's filter-style blog, curating information and news about the team and posting it, generally with significant commentary. However, while many of WiiM's posts either revolve around links to other content or involve original reporting or analysis, some posts are focused more on expression of the author's personal thoughts and opinions, still topic-focused but similar in purpose to diary/journal blogs. In this sense, WiiM shows some hybridity in terms of categorization.

Chapter 3

Situational Characteristics of Winging It in Motown

1. Introduction

As personal computers and access to the internet have become increasingly common, the last three decades have seen technological advances at an unmatched rate. A variety of communicative genres have sprung up as internet-based media, one of the most diverse of which is the weblog, or blog for short. As early blog scholar Rebecca Blood defined it, a blog is a “frequently updated Web site, with posts arranged in reverse chronological order, so new entries are always on top” (2003). The proliferation and popularity of blogs combined with their diversity of both form and topic makes them an important topic of research, and register analysis of a multitude of different types of blogs will go a long way to advancing understanding of blog text from a variety of perspectives. Register analysis of blog text can be useful for everything from marketing to forensic linguistic examination. Garden (2012) suggested that it is crucial for researchers to “provide clear and unambiguous definitions [of blogs] appropriate for their particular research” (p. 483). A register analysis of the blog or blogs on which research is being carried out is crucial to developing this important definition.

The purpose of the work carried out in this chapter is to develop a baseline of characteristics of a specific blog, *Winging It in Motown*, as a text type for use in the authorship identification work that will be carried out later in this dissertation utilizing authors and posts from the same blog. In section 2 of this chapter, I present a discussion of the early history of weblogs and their development. In section 3, I cover frequently-cited characteristics of blogs. In section 4, I discuss Biber and

Conrad's register analysis as it acts as a framework for this analysis on WiiM's posts. In section 5, I cover basic characteristics of Winging It in Motown as they relate to the form of the blog. In section 6, I apply Biber and Conrad's situational characteristics, as derived from Biber & Conrad (2009), to WiiM as a text type. In section 7, I discuss the usage of tagging and counting programs on the corpus created of WiiM posts and analyze the results from the standpoint of linguistic features. In section 8, I discuss the results of running the WiiM corpus through Biber's MAT analysis (Nini, 2014) as a further dimension of this analysis. The MAT analysis aids in further situating WiiM among other registers, showing to which registers the blog is similar and in what ways, as well as which registers are more dissimilar.

2. Background Information on Blogs

2.1 Early History of Blogs

The first step to discussing blogs is to look at their early history, presented here as a brief summary of chapter 2. Per Rebecca Blood (2000), the term *weblog* was coined in 1997 by Jorn Barger. Early blogs were "rudimentary in design and content" according to Blood (2004, p. 54). In early 1999, there were a very small number of blog pages on the web. Over the course of that year, an innovation shifted the trajectory for this register of language use: several new online platforms, including Pitas and Blogger, were developed that would offer a simple way for individuals without strong computer coding skills to create and publish their own blog pages. The development of these software options geared toward non-experts led to an explosion of blog pages on the web, which conservatively numbered somewhere between three and five million by 2005 (Crystal, 2009).

2.2 Frequently-Cited Characteristics of a Blog

What constitutes a blog is a difficult question to answer. With the wide variety of blogs available on the internet, there is great variation among the descriptions of individual blogs. There are, however, some characteristics that are common to most. Researchers frequently cite listings being displayed in reverse chronological order as a common characteristic among most blogs (Herring, Scheidt, Wright, & Bonus, 2005; Bar-Ilan, 2005; Schmidt, 2007; Brala, 2008; de Moor & Efimova, 2004). According to Blood (2000), "the original weblogs were link-driven sites" and were made up of "a mixture in unique proportions of links, commentary, and personal thoughts and essays." Blogs in existence around the time of Blood's earlier work were largely created for the curation of web content in the form of links, sometimes along with blogger commentary. According to Crystal (2009), "links are very important" as a defining feature of the weblog. Thus, links became a commonly-cited characteristic to make a blog as such, and per Crystal, "some blogs consist of little more than a long list of hyperlinks" (pp. 240). Page, Barton, Unger, and Zappavigna (2014) all classified blogs as social media. According to Schmidt (2007) blogs can be multimodal, including text, images, and sound files. Nowson, Oberlander, and Gill (2005) noted that blogs "contains news and views on a variety of topics" and "are already seen as a powerful news-gathering medium" (pp. 1666).

3. Biber and Conrad's Register Analysis as a Framework

I utilize Biber and Conrad's (2009) register analysis as a framework for the examination of a publicly-accessible hockey blog, *Winging It in Motown*, found at www.wingingitinmotown.com. Biber and Conrad's analytical framework involves two major branches of analysis: the description of situational characteristics of the text

and analysis of the text's linguistic features as well as their functions. Their (2001) definition of register is "a cover term for any language variety defined in terms of a particular constellation of situational characteristics" (pp. 3). They also note that "there are usually important linguistic differences across registers that correspond to the differences in situational characteristics" (pp. 3). Biber et al (1999) delves more deeply into the occurrence and analysis of individual lexical and function word categories in texts. Biber developed his Multidimensional Analysis Tagger (Nini, 2014) to compare the rate of occurrence of a wide variety of words and phrases evidencing numerous grammatical categories across text types and group those text types into categories he created, called dimensions. According to Biber and Conrad (2001), "register analyses of these core linguistic features are necessarily quantitative, to determine the relative distribution of linguistic features," and "such analyses require a comparative approach" (pp. 5). The results of this register analysis provide a baseline for eventual authorship studies by outlining which linguistic features may be common to the register, driven by their function as relates to that register, rather than occurring as part of author idiolect.

4. Form

The first part of Biber and Conrad's binary approach is an overview of the form structure of the text. An understanding of these characteristics is crucial to drawing connections between the formatting of the text, its situational characteristics, and its linguistic features- that is, discovering potential functions for those features as relates to the blog as a register.

4.1 Examining a blog's form

The blog used for this analysis, *Winging It in Motown*, is a public blog that is

easily accessible through a straightforward web address: winginginmotown.com. Furthermore, I, the researcher, have a long-standing membership with the blog and am familiar with its bloggers, form, and content from years of use as well as previous ethnographic study. This allowed me to cover each aspect of Biber and Conrad's analysis of the form of the text thoroughly. For the purposes of this discussion, the form of the blog as it existed at the end of the process of developing the corpus, in March of 2017, is the only iteration of the blog that was examined and will be discussed. As the blog has undergone some significant visual and user interface changes over the last two years, it is important to note the point of time at which the form of the blog was examined. I first closely examined the home page, which has evolved from quite simple and straightforward to rather complex over the years. This was a much more involved process than during previous ethnographic study, and because the home page has become more complex, the discussion of its form has as well. Next, I explored the entirety of the site map and examined individual blog posts for common form characteristics. Some aspects of the discussion of form require an understanding of the content of main blog posts, fanshots, fanposts, and comments, so I read through a variety of representative posts and comments from multiple topic sections as well as gaining insight from my existing experience as a reader of the blog. These examinations and experiences allowed me to analyze the form of the blog thoroughly from each angle discussed by Biber and Conrad for their register analysis procedure.

4.2 Post frequency

Per Garden (2012), "frequent updates are... considered important" in defining a website as a blog. WiiM is a prolific blog. As of March 31st, 2017, WiiM saw 344 blog posts, and there were 1293 posts in 2016 and 1269 in 2015, the years from

which the corpus data was culled. Based on data from 2016, that means an average of almost 108 posts per month, almost 25 per week, and over three and a half per day. While one must take into account that this is a sports blog, and sports produce times during the off-season when there may not be much to discuss, even in August of 2016- the month generally seen as the slowest in NHL news- saw 49 blog posts. These counts include only posts that were posted to the blog by the official bloggers. Fanshots and fanposts are not included. Bar-Ilan (2005) conducted research examining 15 weblogs over 61 days for aspects including posting frequency and found post-per-day averages ranging from .11-4.85, with the average for all 15 blogs at about 1.17 posts per day. Compared to these results, WiiM's over three-and-a-half-per-day average places the blog on the prolific side of the spectrum.

4.3 Home Page

The formatting and appearance of WiiM's home page have changed significantly since this research began, and the home page is now much denser than previous iterations. The page is now topped with a paid

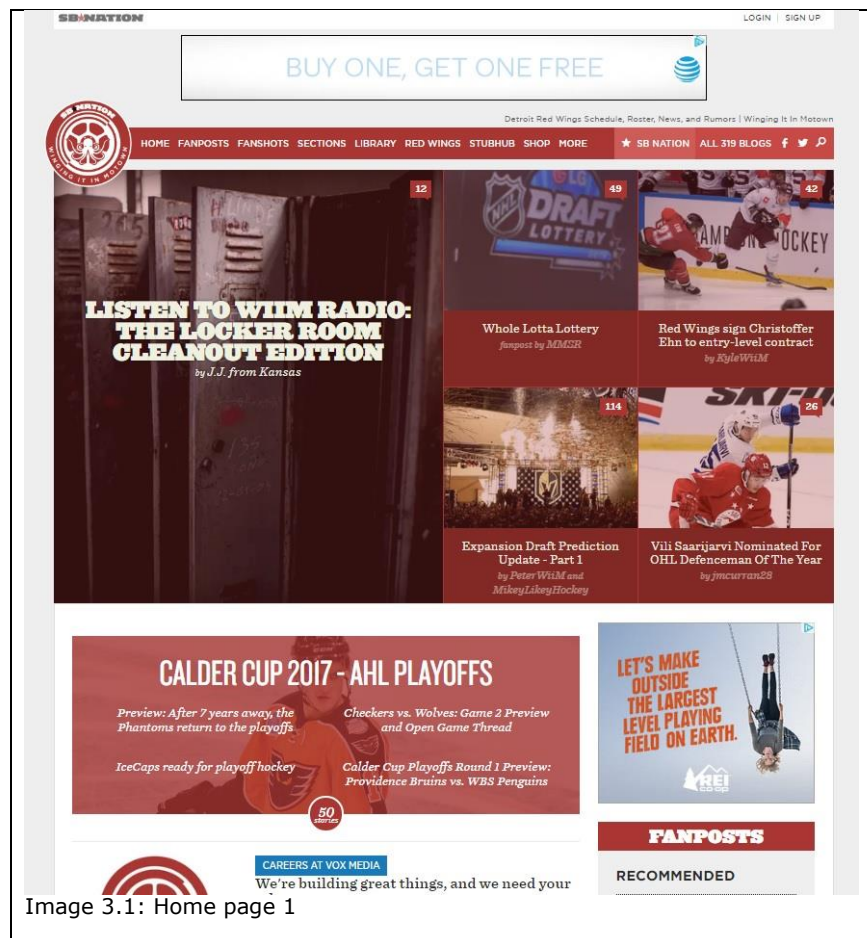


Image 3.1: Home page 1

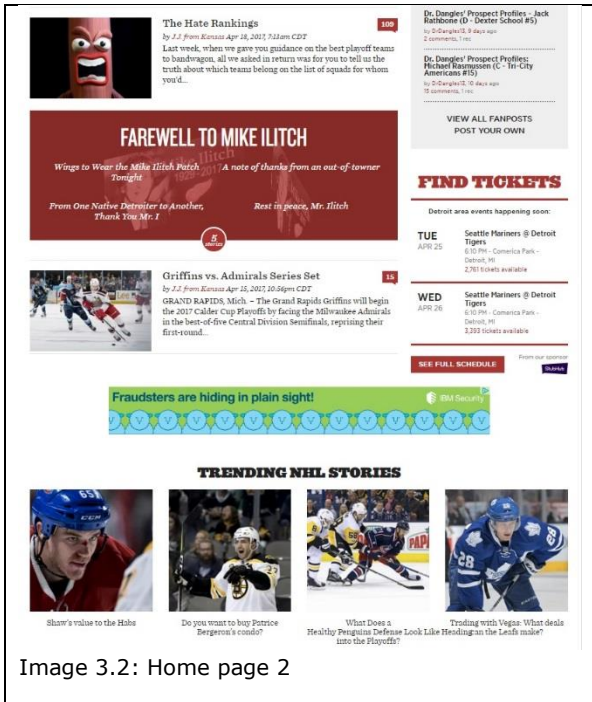


Image 3.2: Home page 2

advertisement, often large enough to take up a significant portion of the viewing screen. This advertisement changes frequently and often surrounds the top portions of the images and text of the blog itself. Surrounded by these ads is a navigation bar of button links with WiiM’s customized logo. These buttons include: *Home*; *Fanposts*; *Fanshots*; *Sections*; *Library*; *Red Wings*; *Stubhub*; *Shop*; and *More*. There is

also an SB Nation vs. Admirals button which changes the bar to show the variety of sports and sports leagues for which SB Nation hosts blogs.

The bar also has a button that links to an SB Nation directory page for all 319 blogs, quickbuttons to allow the user to Facebook-like or Twitter-follow WiiM directly from the bar on the homepage, and a search button. Five featured posts take up the rest of the viewing screen. These posts are presented as images which link to the posts directly. The posts’ titles are shown as well as their authors, with the post’s headline photo as background to the text. The primary featured post is shown in a larger box, with the other four together in smaller boxes next to it. Each of these boxes also includes a quick link button to the comments which displays as the present number of comments for that post.

Under these featured stories is a graphic to showcase live game scores, with boxes to break down goals by period per team. Scrolling down the home page past

these images brings the viewer to a longer list of posts, which appear to primarily be in chronological order beginning with the newest, but which have promoted or pinned posts interspersed. Also showcased among these posts are boxes which cluster together posts that are relevant to each other or follow a specific topic or storyline, featuring a title for the box and titled links to four relevant posts. Along the side of this list of new posts is a separate section which lists and links to recent fan posts, which are kept separate from posts authored by official WiiM bloggers. Below the fan posts list is a specialized box labeled *Find Tickets* which showcases upcoming Red Wings games and links to ticket information via the ticketing website StubHub. While this box has an element of advertisement, it is specifically designed to help readers quickly access tickets to specific upcoming games.

The box does not change to feature advertisements for other products or services as the advertisements at the top of the page do. As the viewer continues to scroll, another large advertisement box is encountered, followed by a box titled

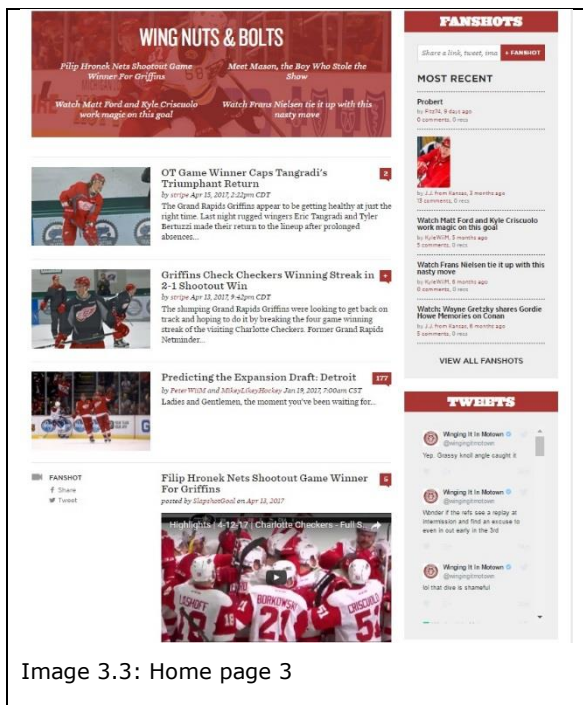


Image 3.3: Home page 3

Trending NHL Stories which features images and links to posts on SB Nation blogs which follow other teams within the league. Below this box the chronological list of WiiM posts continues, this time with a side box featuring Fanshots, non-WiiM internet links submitted by fans as relevant to the topic of the blog. Below the Fanshots side box is another side box showcasing recent tweets from WiiM's

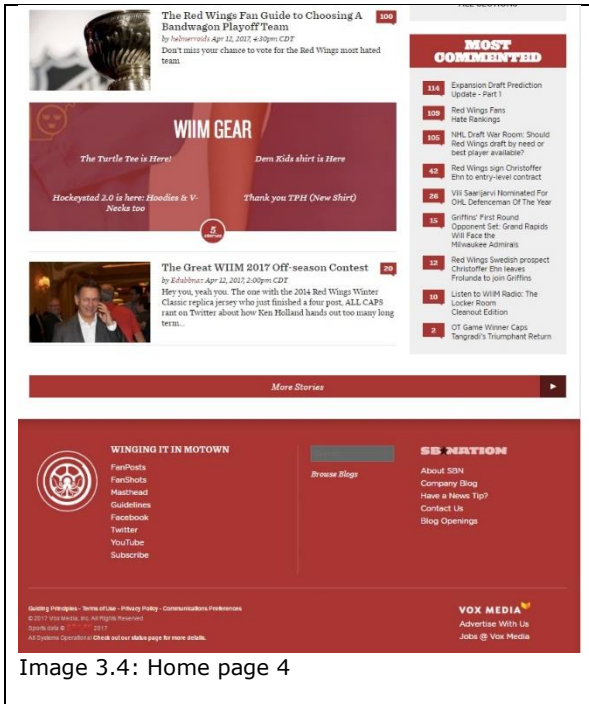


Image 3.4: Home page 4

Twitter account and another advertisement box. Another interruption to the chronological list makes an appearance, a box labeled *Popular Topics* which showcases current trending sports topics, each of which links to a list of SB Nation posts relevant to that topic. Below this, the chronological post list continues, with new side boxes: a *Featured Sections* side box showcasing stories from that day's featured blog sections, and a

Most Commented box, which includes a list of links to the most-commented-on recent posts in descending order and the number of current comments. Below this side box and the chronological posts list is a bar-style link button titled *More Stories* which takes the viewer to the next page of posts listed chronologically. Following the *More Stories* bar is a box with links and information about WiiM and SB Nation. The WiiM section includes links titled *Fanposts*, *Fanshots*, *Masthead*, *Guidelines*, *Facebook*, *Twitter*, *YouTube*, and *Subscribe*. The SB Nation section includes a search box and a link to browse all SB Nation blogs, as well as links titled *About SB Nation*, *Company Blog*, *Have a News Tip?*, *Contact Us*, and *Blog Openings*.

4.4 Beyond the Home Page

The *Sections*, *Library*, *Red Wings*, and *More* buttons all open drop-down menus with further, more specific link options available. All further pages also feature ads along the sidebars and below the main portion of the page. *Sections* includes

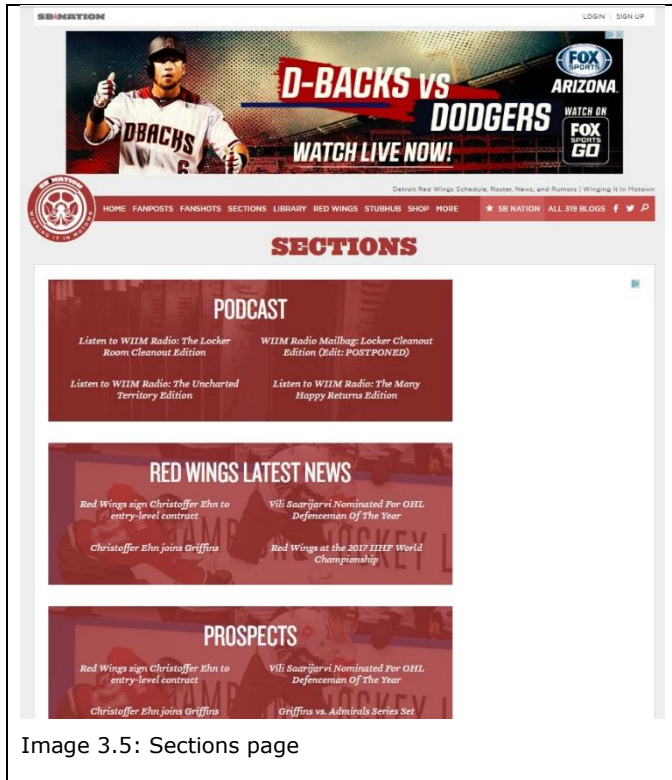


Image 3.5: Sections page

quicklinks to posts about Red Wings prospects, podcast posts (which include links to the podcasts), posts regarding the death of Red Wings owner Mike Ilitch, a post series entitled *Getting to Know Advanced Stats*, and one entitled *Getting to Know the CBA* (Collective Bargaining Agreement), as well as a quicklink to the full post archive. Clicking on the *Sections* button instead of hovering over the

drop-down menu takes the user to a page with a variety of quick-links sorting the blog posts by topic, including *Opinion*, the farm team *Grand Rapids Griffins*, *Game Threads*, and *Quick Posts*. There are 47 total topic-based sections on this page.

The *Library* button is not a quicklink in itself, but offers three quicklinks in a dropdown menu: *Reference Links & Documents*; *WIIM's Getting to Know Series*; and *Blogroll*. The *Red Wings* button, like the *Sections* button, offers both a dropdown menu and a quicklink. The dropdown menu is the longest of the bar buttons, offering links to *Stories*, *Schedule*, *Roster*, *Stats*, *Yahoo Red Wings News*, *Yahoo Red Wings Team Page*, *Yahoo Red Wings Report*, *Yahoo Red Wings Depth Chart*, *Yahoo Red Wings Transactions*, and *Yahoo Red Wings Photos*. The quicklink brings the user to a page with a large box at the top showcasing the Wings' current season record, the score box from the last game, an information box for the next game with a link to

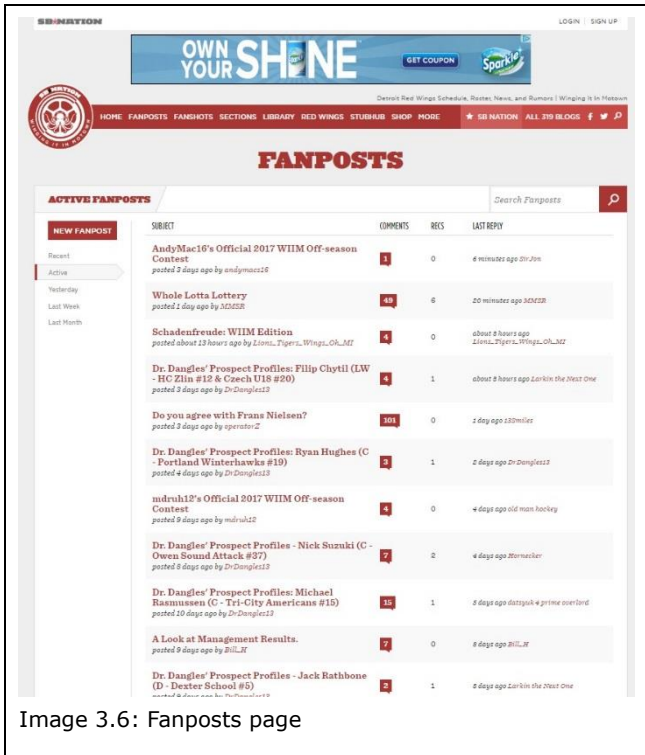


Image 3.6: Fanposts page

buy tickets, and a dropdown box to view a similar page focused on an individual player. Below this large box is a chronological list of stories from all SB Nation blogs that are tagged as discussing the Red Wings, as well as options to replace this section with schedule information or roster information instead. The *More* button, like the *Library* button, is not a quicklink itself but offers a small dropdown

menu. This menu includes a link for *Odds* for betting information and a link titled *About*, which shows profile links and information about each WiiM author.

The buttons which do not offer dropdown options- *Home*, *Fanposts*, *Fanshots*, *Stubhub*, and *Shop*- are all direct quicklinks. The *Home* button simply takes the user to the primary homepage. The *Stubhub* button takes the user out of WiiM and SB Nation entirely and directs them to the Red Wings page of the ticket retailer Stubhub, where the user can

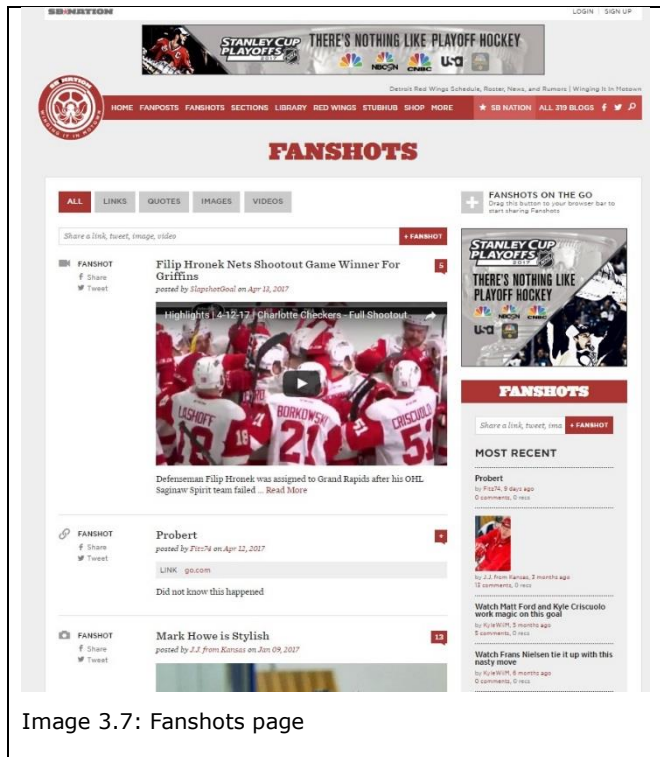


Image 3.7: Fanshots page

purchase tickets to an upcoming game. The *Shop* button also takes the user out of WiiM and SB Nation, in this case directing them to the online sports merchandise retailer Fanatics, again straight to their Red Wings page. *Fanposts* and *Fanshots* are sections that are particular to SB Nation and are not as self-explanatory as other button links. Fanposts allow a space for users who are members of SB Nation but not official bloggers to post their own thoughts, information, and so on as blog posts. This section is kept entirely separate from official posts by the designated authors, but the design features of the posts are very similar, including a comments section to allow interactivity. Fanshots are a similar concept, but instead of allowing users to author their own posts, this section is for links which the users wish to share as relevant to the blog and its readers. These links appear as their own post, with no additional text authored by the posting user. Both of these sections require users to be logged into their SB Nation profile in order to post.

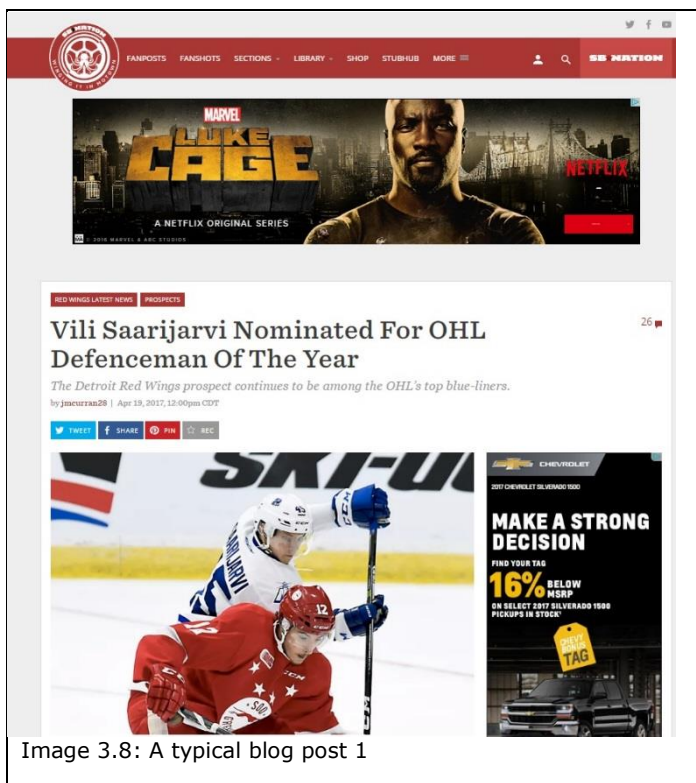


Image 3.8: A typical blog post 1

A. The Design of a Blog Post

If the user navigates to a specific blog post, there is a common format that they will encounter. Gone is the large ad found at the top of the home page, but the navigation bar with its previously-described buttons remains, along with the WiiM logo. A small ad is shown below the bar, and below that

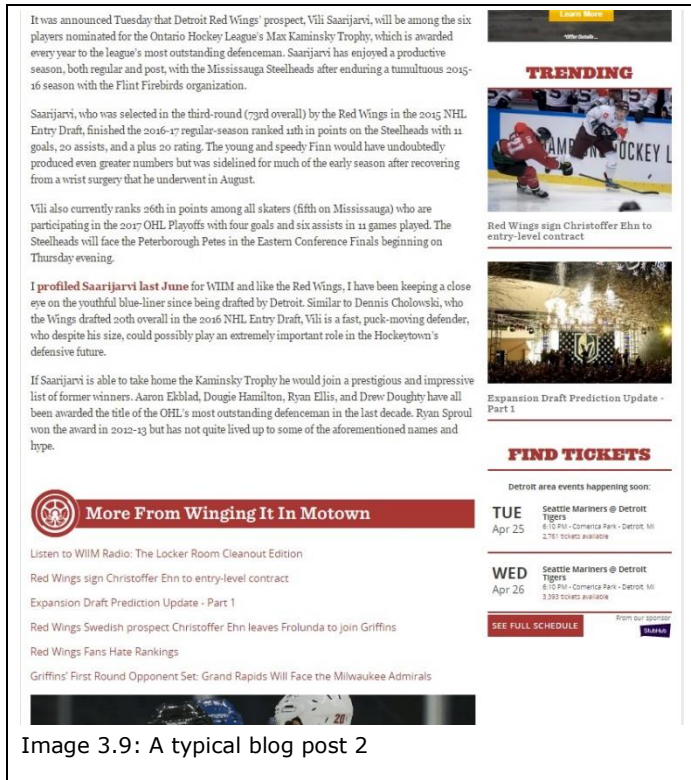


Image 3.9: A typical blog post 2

ad is the blog post box. At the top of the box is a graphic indicating the section in which the post belongs, which is both a graphic and a quicklink to the *Sections* page. Below this is the title of the post, along with a comment count graphic which is also a quicklink to the comments on the post. The author byline is below the title, showing the SB Nation username of the author, which

is a quicklink to his or her profile, the author's Twitter handle, which is a quicklink to their Twitter page, and the date- and timestamps for the post. This information is followed by buttons to share the post on Twitter, Facebook, or Pinterest, and a *Rec* button- an SB Nation-specific feature that allows users to "recommend" the post much in the same way a user utilizes the *Like* buttons on Facebook and Twitter. Below these buttons is generally an image representative of the story and a credit byline for the image. Below the image is the story itself, sometimes with ads interspersed. After the story is a banner titled *More From Winging It in Motown*, under which can be found links to other recent stories posted to the site. Presently, this section of the posts is followed by a video about a day in the life of an NHL referee, and then more advertisements. Below the ads is a *Recommended* section with both sponsored links and links to other blog posts. Finally, as the final section of the blog post page, the user will find the comments section. A user must be logged

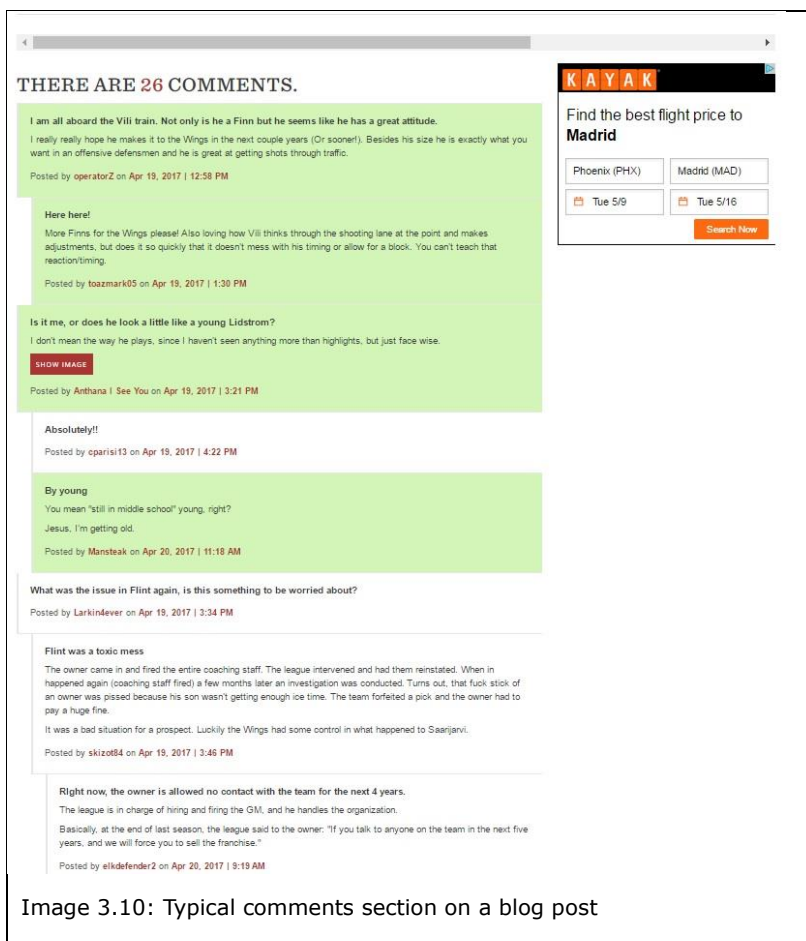


Image 3.10: Typical comments section on a blog post

in to use this section. The blog posts feature a side bar similar to that found on the home page, containing advertisements, a *Trending* section of popular posts, a section with links to tickets for upcoming games via StubHub, and a *Team Shop* link showing several merchandise items which link to those

items for sale on the Fanatics website.

5. Situational Characteristics

Defining a text's situational characteristics is an important part of describing the text as a register. In the case of WiiM, these characteristics may vary between individual blog posts, depending on the subtype of the post in question. A typological examination of individual WiiM posts is beyond the scope of this paper, but an overview of the common situational characteristics across the blog as well as a brief discussion of their variances from post to post will aid in situating linguistic features as characteristic of the register rather than e.g. author idiolect. This determination will eventually aid in determining best characteristics for examining idiolect-driven

language, in this case for the purpose of authorship determination.

5.1 Participants

The first aspect of the participants category that must be discussed is that of the addressors of the communication. While many blogs claim only a single author, WiiM has three main bloggers who also act in an administrative capacity, referred to as *The Managers* on the *About* page, as well as four primary contributors who

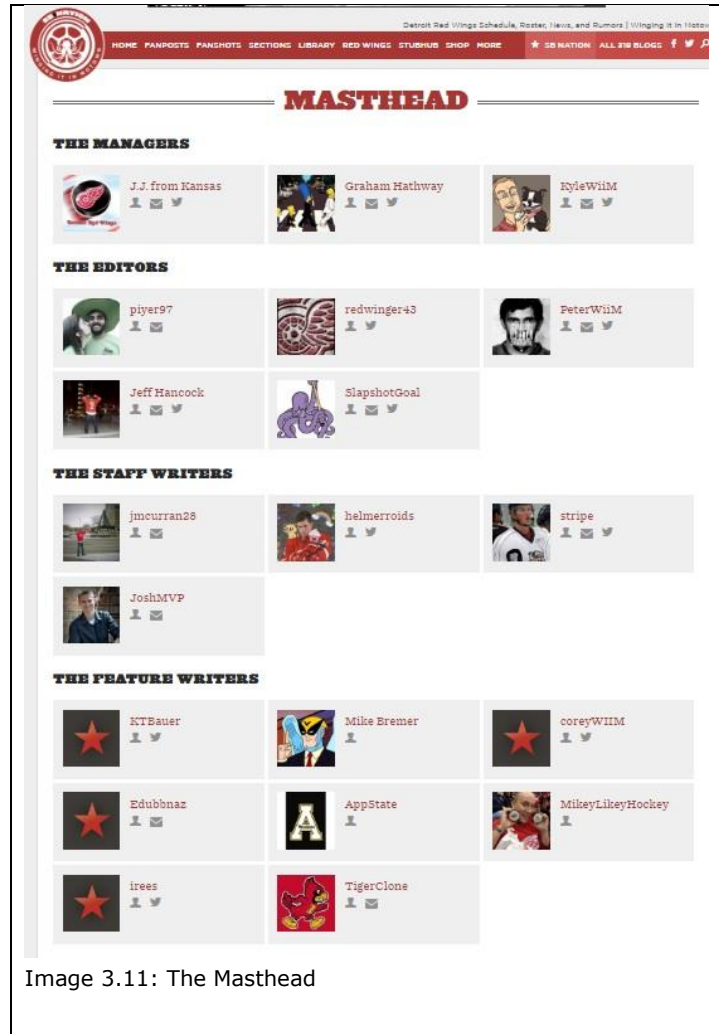


Image 3.11: The Masthead

have regular blog post series, *The Editors*, four more contributors who frequently author posts, *The Staff Writers*, and eight contributors who are authorized to post regular blog posts but do not post frequently or only cover very specific topics, *The Feature Writers*. With a grand total of 19 potential authors at the present time, WiiM definitely qualifies as falling under the plural category for addressor characteristics. Having this many post authors on one blog is an unusual feature, and furthering that is the fact that WiiM has had additional authors in the past, whose blog posts remain on the site but who no longer contribute as official authors. Thus, taking the entire blog, with all of its post history, into account, the list of authors is, in fact, even

longer than 19. WiiM does not have an institutional addressor- the posts are attributed to individual authors via the author byline, not to an institution. The authors of WiiM posts also do not qualify in the category of unidentified. Not only are they identified in the author byline by their SB Nation username, but that username is directly linked to an informational profile on the *About* page. These profiles include post and comment counts and a list of direct links to recent activity, both in terms of comments and in terms of posts, e.g. one can click on a recent comment from the author and be taken directly to that comment and the page it resides on. The profiles also include the date the individual joined, a search box to search that user's specific activity, and a brief biographical blurb written by the author. Most importantly in terms of identifiability, the profiles include links to the author's web page, social media accounts, and email address, offering a way to at least identify the individual on other platforms. While the posts of these official bloggers are kept separate from Fanposts and Fanshots, both of those categories require users to be logged into their SB Nation accounts, and they still include author bylines on posts. Those bylines again link to profiles with the ability to include the same information as available on the bloggers' profile pages, though the user-added information is not required. This allows all contributors of posts to this site to be identified at least at the username level.

WiiM sees a variety of social characteristics among its contributors. This information is not readily accessible on the site, via the profiles or on any other structural aspect of the blog. However, WiiM is a fairly interactive site, offering several ways for users to interact with bloggers, and the links to personal webpages, social media accounts, and email address on the profile pages allow for still more interactivity. Many of the bloggers are deeply involved in the wider online Red Wings

community and thus interact frequently with readers in other corners of the internet besides the comment sections of WiiM. Many of them are particularly active on Twitter and encourage discussions with readers in that arena. This allows for additional ways to discover their social characteristics. A new reader of the blog would not necessarily know any social characteristics of the bloggers, but someone who frequently reads the blog, interacts in comments, and interacts with the authors on Twitter is likely to figure out some of this information. Several of the feature authors included no social media links and have not contributed frequently in terms of comments or posts, making it very difficult to identify any of their social characteristics. Three authors could be positively identified as female. Three could not be identified in terms of gender. The other 13 are male. Ages vary, but the majority of authors for whom these characteristics were discoverable were between young adulthood and middle age. Similar variation exists for other characteristics. Some bloggers are full-time workers in other careers or industries, while some are students. Their locations are spread out, some residing thousands of miles from Detroit, the location of the team the blog is centered on. Of the three head bloggers, *The Managers*, only one lives in Michigan, where both the Wings and their farm team affiliate are located, while the other two reside in Kansas and Illinois.

The other side of the participants of this interaction, the addressees, must be described as well. The intended audience of a communicative act can have a direct impact on the decisions the actor makes while executing the act. This is particularly true where concerns written communication, and even more so in asynchronous communication such as blog posts. As blogs are on the internet where they can be accessed by virtually anyone as long as they have internet access, unless the blog is deliberately password-protected or otherwise secured to control who has access, the

writing is typically driven by the idea that anyone who comes across the post and wants to read it is who the post is written for. This holds true for WiiM. Although logging into an SB Nation profile is required for posting Fanshots and Fanposts as well as for commenting, the site has no such restrictions for access to the posts themselves. If one wishes to read a post, one must merely navigate to the page on their browser, which will allow them to explore all previously discussed features and read any and all blog posts, fanposts, and fanshots. Comments can also be viewed regardless of log in status. As long one is not actually trying to contribute, the user has access to every corner of the site. Clearly, WiiM does not have a single addressee. However, the site does not necessarily have a plural addressee either. While writing the posts, it is not possible to identify with assurance the set of addressees- that is, the blogger cannot imagine specifically who will be reading the post, no matter how many readers there may be. It is not even possible for the blogger to imagine with certainty how many readers there could end up being for that specific post. It could be one or thousands, and any person on the internet could be among those numbers. Thus, WiiM's addressee falls under the unenumerated category. This idea the blogger must keep in mind, that anyone from anywhere at any time may consume their post, may have a significant impact on the linguistic choices the blogger makes, both consciously and subconsciously, and this impacts the overall language of the blog as a register.

The final angle from which language must be examined in terms of participants involves the concept of onlookers. WiiM is overall an asynchronous form of communication, and this is entirely the case where concerns blog posts specifically. The comments sections can be used in a capacity that is similar to a live discussion, particularly in the posts marked as game threads, which are designed

specifically for this interactive activity. In that case, there may be onlookers to a discussion occurring between two or more posters that becomes something akin to a synchronous conversation who may never engage in the conversation themselves. However, this scenario is not relevant to the blog posts themselves. In the case of the posts, anyone engaging with the post is either a writer or a reader, and thus either an addressor or an addressee, at any given time. Anyone who is not the author of the post but is reading the post falls under the umbrella of the unenumerated addressees of the post.

5.2 Relations among participants

The second dimension of Biber and Conrad's register analysis technique is an examination of the relations among participants. The first characteristic to cover is the interactiveness of the register, which has already been touched on several times in this discussion in regard to WiiM. While blog writing is generally an asynchronous activity designed as a product produced for others to consume rather than as a dialogue-style communicative interaction, many blogs include comment sections that introduce an element of interactivity. These comment



Image 3.12: Blogger JJ interacting in the comments

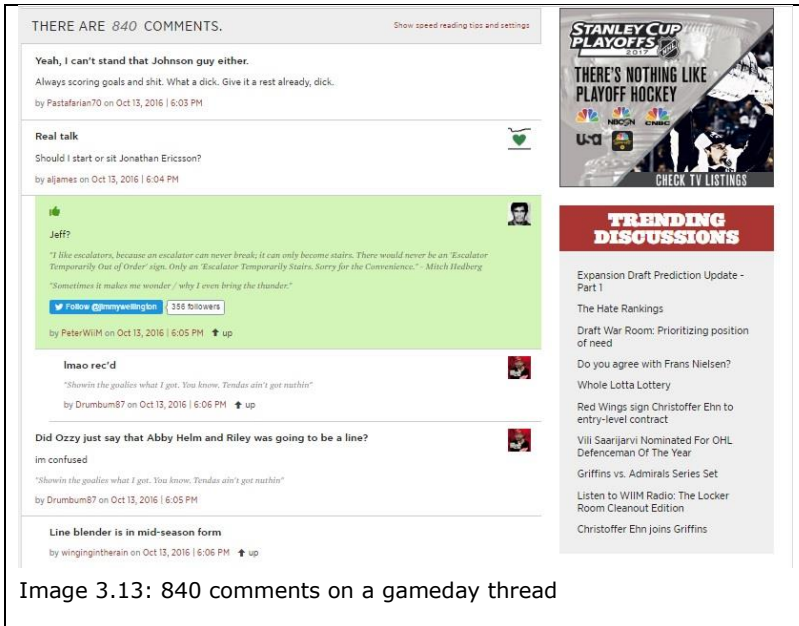


Image 3.13: 840 comments on a gameday thread

sections allow readers to react to the post in a way that allows the author to see what they have to say. As authors can also post comments, this also allows them to respond to the reader's thoughts and continue

a dialogue, potentially indefinitely. WiiM is a highly interactive blog. Every post has a comments section open and available to all logged-in readers, and comment counts are typically quite high. Bar-Ilan's (2005) 61-day survey of a collection of blogs found that not all blogs had comments enabled even when the option was available. She also discovered that the most comments received during the entire 61-day period was 369, by the blog *Online*, around 5% of the number of comments received by WiiM posts in just March of 2017 with gameday threads designed specifically for live discussion-style commenting excluded. These gameday threads generally see between 200 and 800 comments per post. The highest comments per post Bar-Ilan found on any of the blogs she studied was

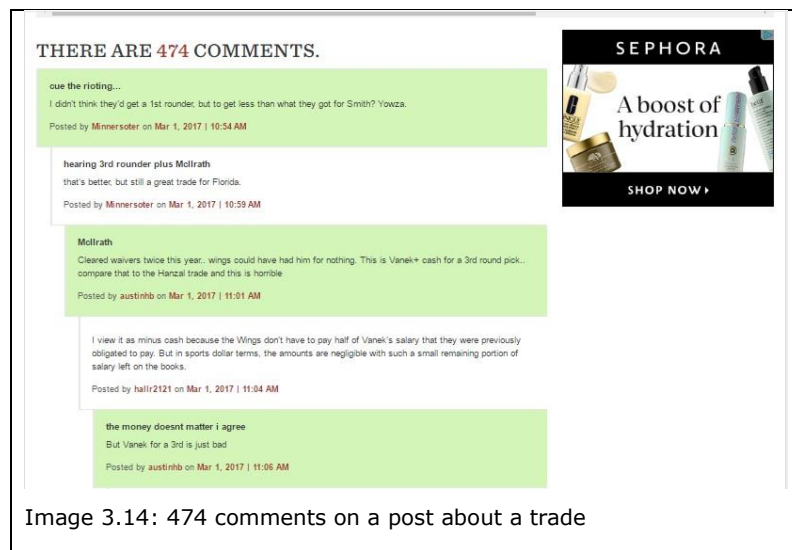


Image 3.14: 474 comments on a post about a trade

23. WiiM's gameday thread for the first game of the regular 2016-2017 season had 840. A March 1st, 2017 post about the trade of a Wings player saw 474 comments. This suggests a solid placement for WiiM as a highly interactive blog.

WiiM has several different features which promote its interactivensess in addition to merely allowing comments on all posts. The aforementioned gamethreads provide the primary avenue for interaction. Each day there is a Red Wings game, one of the bloggers creates a post specifically for discussion live during the game. The blog post itself is very sparse, with little or no actual commentary from the blogger. The post contains a score box which is updated live throughout the game. There is also an informational box which tells who the team is playing, each team's record as of that day, each team's starting goaltender, the corresponding SB Nation blog as a quicklink for the opposing team, game time and television stations broadcasting, and, as a special feature for this year, a countdown of how many games the team has left to play at their present arena, which is soon to be replaced. These features are standard in the game day posts and make up the bulk or the entirety of the post.

The primary purpose is really to allow an interactive place for people to discuss the game as it occurs, in the form of the

Gamethread: Red Wings vs Lightning
 By J.J. from Kansas @JJfromKansas on Mar 24, 2017, 6:00pm CDT 365

Final

Team	TOTAL
Tampa Bay Lightning	2
Detroit Red Wings	1

Saturday, March 24, 2017 - 1:20AM EST Standard Time

Tampa Bay Lightning	Team	Detroit Red Wings
25-28-9 70 points	Record	29-22-11 69 points
Andrei Vasilevsky	Starting Goaltender	Petr Mrazek
Raw Charge	SB Nation Blog / tail guy	Sustr
7:30 PM EST	Time - Television	FSD, SUN

Let's do fun stuff.
 Have fun in the comments and let's go Red Wings.

CRUISE.COM
 Alaska Cruises
 From \$549
 View

LATEST NEWS
 Listen to WiiM Radio: The Locker Room Cleanout Edition
 Whole Lotta Lottery
 Red Wings sign Christoffer Ehn to entry-level contract
 Expansion Draft Prediction Update - Part 1
 Ville Saarijani Nominated For OHL Defenceman Of The Year
 The Next 18 Months - The Red Wings' Road Back - Part 3

Image 3.15: A typical gamethread

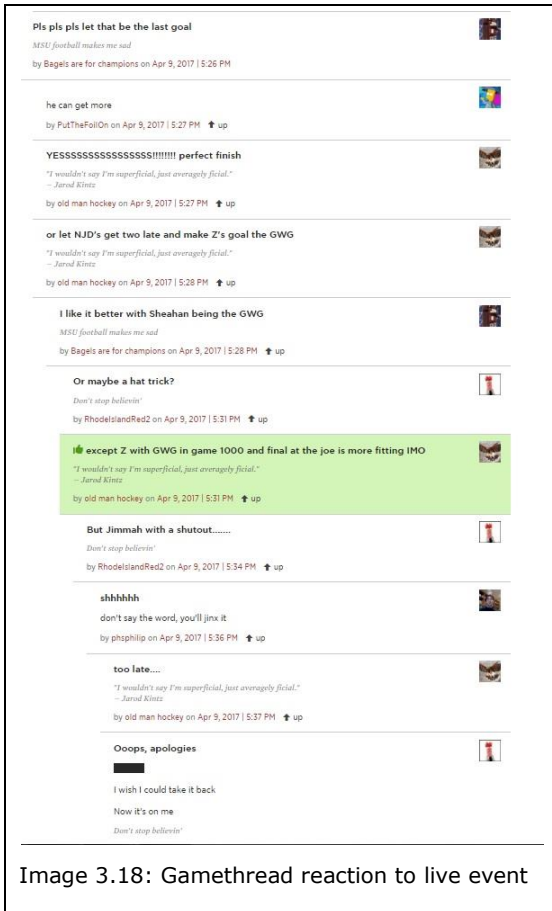


Image 3.18: Gamethread reaction to live event

season and Sheahan’s difficulty scoring to understand. Participants also frequently reply rapidly to each other’s posts, with the speed seen in messenger programs such as Google Talk or Yahoo Messenger rather than the typical delay in replying to comments on many blogs. Time stamps on comments in these blog posts show multiple posts and replies occurring within a matter of minutes. Furthering the argument that this

additional contextual information. Readers unaware that Riley Sheahan had just scored during the live game would be confused enough, but those unaware of the additional context that Sheahan was scoring his first goal of the season in the final game of the season would be even more confused. This collection of comments requires both the contextual information of the live game and the contextual information of the entire

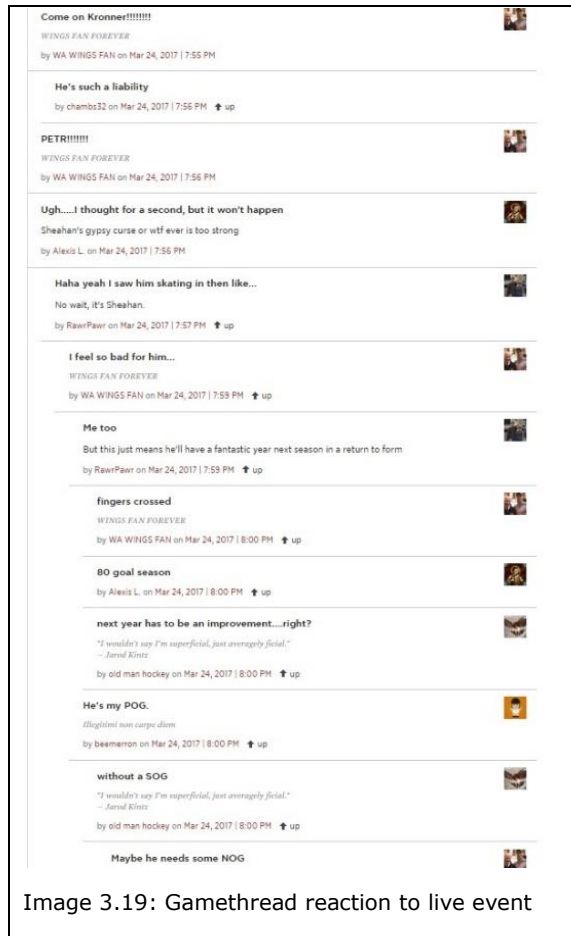


Image 3.19: Gamethread reaction to live event

promotes a high level of interactivity is the frequent and strong presence of many of the bloggers in these conversations, as they are often not only commenting but heavily involved in the rapid discussions that take place.

The previously discussed Fanpost and Fanshot sections of the blog contribute the final two features which elevate the interactivity of WiiM. The Fanshots section allows the user to quickly and easily share a link with the community that they deem relevant, without requiring any commentary on their part. The purpose of this option is not as a place for users to editorialize, but rather to provide a quick and straightforward way to directly share information obtained elsewhere on the web with the community at large. SB Nation users can share photos, videos, articles, and a variety of other types of links directly via this mechanism that perhaps have not been shared in a Quick Hits post by the bloggers and thus may otherwise not have been seen by members of the community. The bloggers themselves even sometimes

utilize this quick and easy option to share individual links. While fanshots are not designed for users to editorialize, there is also a place for them to do so:

Fanposts. Fanposts provide a platform for users who are not official bloggers to still contribute what are essentially their own blog posts. They are

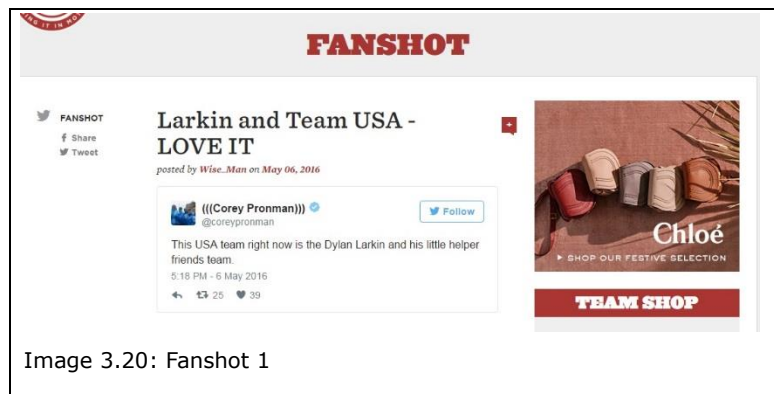


Image 3.20: Fanshot 1

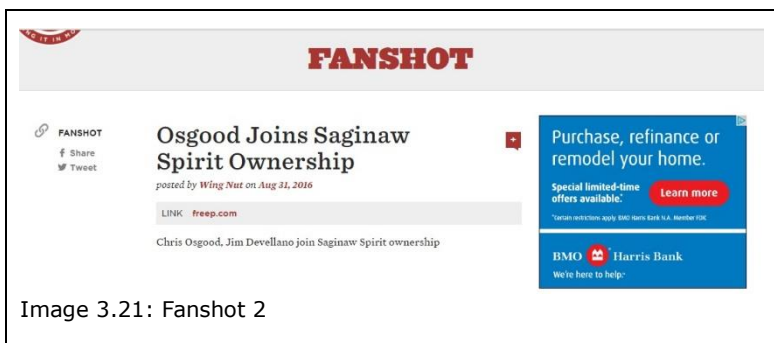


Image 3.21: Fanshot 2

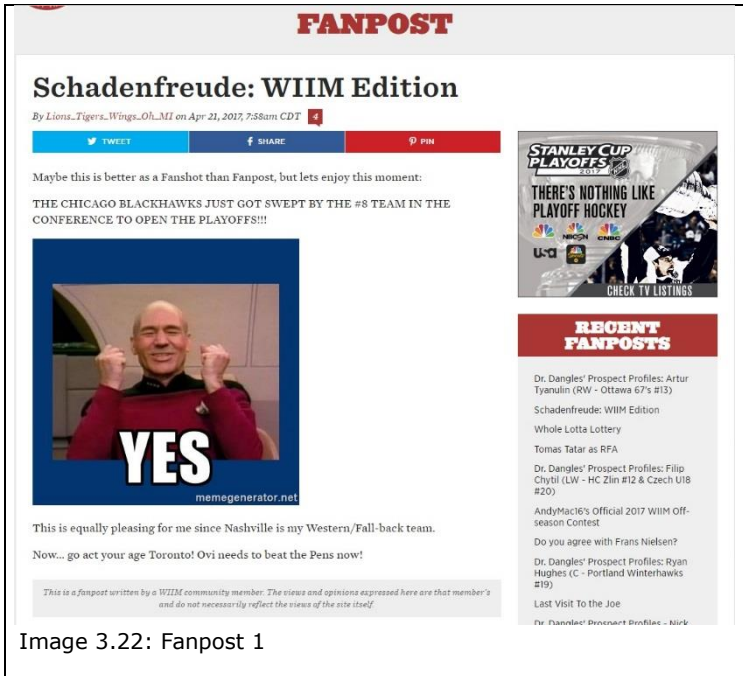


Image 3.22: Fanpost 1

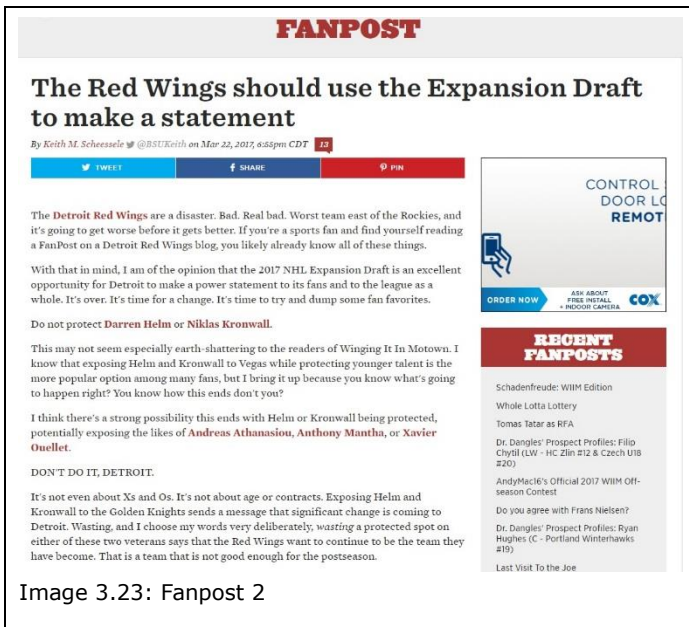


Image 3.23: Fanpost 2

able to analyze and discuss much in the same way that the bloggers often do in their posts, but these are kept in a separate section entirely from the posts of the official bloggers. It is not uncommon for the official bloggers to suggest to commenters that they write up a fanpost based on

a comment on one of the main blog's posts. Both of these sections also include comment sections, allowing for discussions to unfold just as they do in main blog posts, though these posts do often see a lower number of comments overall in comparison.

Social roles are another important component of relations among participants. All of the official bloggers have a higher status and more access and control than other members of SB Nation who function as readers of the blog instead. These individuals have access to the ability to create and edit posts in the main section of

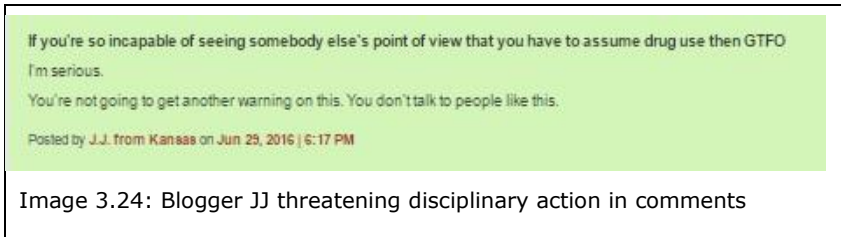


Image 3.24: Blogger JJ threatening disciplinary action in comments

the blog. These posts are shown as the main portion of the blog, rather

than being separated as a special section, and they are permitted to post what and when the please, while Fanposts and Fanshots require administrative approval. These individuals thus have not only a higher level of access, but power to control what is shown on the blog and what contributions from members can become visible to the public. The bloggers are sectioned off, as described in the discussion of the *About* page, and this reflects their hierarchy of power, with *The Managers* having the most access and control over administrative, disciplinary, editing, formatting, and other responsibilities, as well as communication with SB Nation and the handling of issues and complaints from members. They have ultimate decision-making power and the ability to override other bloggers as needed. Some of the bloggers have the ability to edit other content besides their own, access control panels for the blog, and restrict access of other users by suspending or banning their accounts if necessary. They are permitted to act as moderators, and are also the individuals responsible for approving Fanshots and Fanposts. They can also edit or remove comments from comment sections. This administrative control is restricted to the WiiM site and they do not have this level of power on other SB Nation blogs- rather, when they are visiting those blogs, they are mere members just like other readers. They cannot moderate content or alter formatting on other SB Nation blogs, and a ban they have instituted restricts the banned member's use of WiiM only and does not carry over to other blogs. A user banned from WiiM is banned only from WiiM unless the moderators of other blogs ban them from those blogs as well.

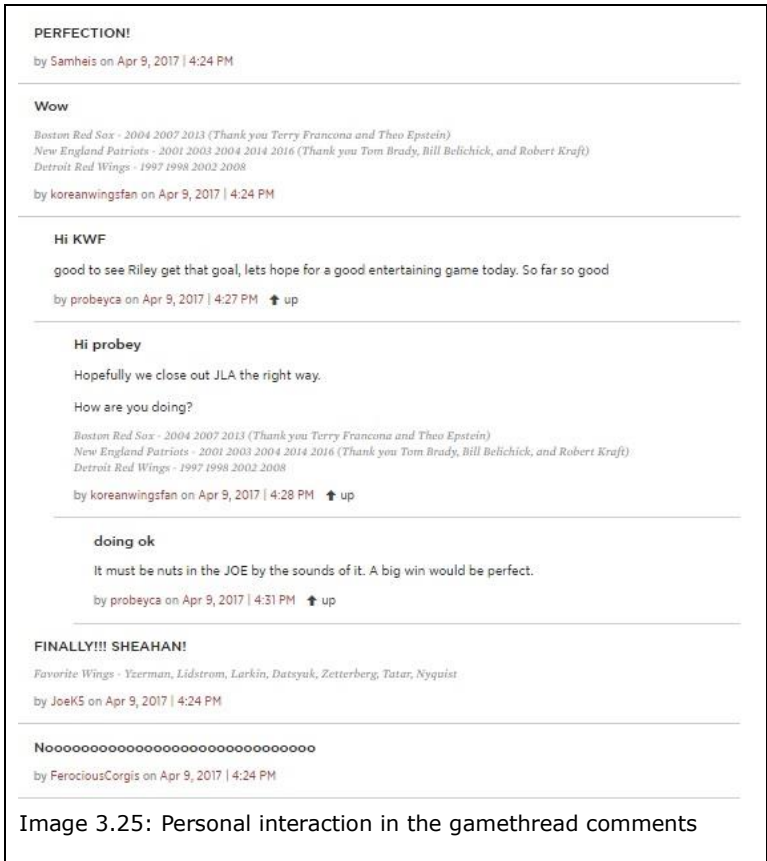


Image 3.25: Personal interaction in the gamethread comments

Another aspect of the relations among participants category that must be discussed is the personal relationships among the participants of the communication. Generally speaking, blog writers who post on blogs that are completely publicly available can assume that the majority of their readers are strangers to them.

Certainly, on a blog like WiiM, with such a large reader base, this is the case. Thus, bloggers will generally write as if their audience is not familiar. However, the online hockey community- and in particular the online Red Wings community- is a rather small world. A number of users on WiiM are familiar with the bloggers outside of interactions on the site and the blogs they post. Many of the bloggers have attended or even organized in-person group meet-ups with open invites to everyone in the WiiM community, leading to some readers having interacted significantly with them in real life. On top of that, many of the bloggers maintain strong social media presences and interact regularly with individuals who are also readers at WiiM on those sites. These interactions can be more personal than the interactions between blogger and reader, including sharing information about personal lives such as work and family experiences. Thus, some readers may be considered acquaintances or

even friends by the bloggers, rather than strangers, and this more personal level of relationship between them and a number of readers can potentially impact their linguistic choices. These differences are notable particularly in the comments sections with bloggers interacting directly with those more familiar users, and less so in the main blogs, which need to be written for the entire audience. None of the bloggers work with each other outside of the blog in their professional jobs. However, the dynamic under which they operate within the blog leads to them having a relationship similar to that of colleagues. In this context, they work together to achieve the same end goal of maintaining a high-traffic blog in such a manner as to see it flourish. Decision-making and troubleshooting issues that arise is something that is carried out as a collaborative effort among the bloggers, and this contributes to a relationship that is similar to a professional colleague.

Shared knowledge is the final element of relations among participants to consider, and this category is examined from two angles: personal and specialist. There is an element of shared personal knowledge in a community such as this. Some of the bloggers are familiar enough with some of the readers to be acquaintances or even friends, some even beyond the realm of the internet. They thus have knowledge of each other's family lives, work situations, and so on, and topics revolving around these more personal elements can enter conversations on the blog, particularly in the comments. However, shared personal knowledge is not expected in this setting, and it is not especially common among participants here. Shared specialist knowledge, though, is certainly expected, and that expectation is heavily relied upon as the context it provides is often presupposed and not given explicitly in the blog posts. WiiM is a topic-driven blog, and the topic is highly specific. The blog is centered on not just sports, not just a specific sport, not just a

specific league, but one certain team, and while the environment is friendly to newcomers who may lack much knowledge on the topic, they will require extra information to be able to fully understand most of the posts. Generally, if a reader needs further information, other readers and bloggers are happy to discuss and provide context in the comments section, but a new reader with little background entering a blog post without the context of this specialized knowledge would likely find themselves overwhelmed and confused. General knowledge of the sport of ice hockey itself is important, as well as knowledge of the team's current roster,

OPINION ANALYSIS/REACTION

Red Wings Mythbusters: Nyquist and Tatar are Bad Together

Are the Wings' two best wingers just bad for each other?

by J.J. from Kansas | @JJfromKansas | Sep 2, 2016, 11:30am CDT

TWEET SHARE PIN REC

Photo by Doug Pensinger/Getty Images

Gustav Nyquist and Tomas Tatar are the Red Wings' two best wingers. If there's doubt about it, then it's likely due to a confusion about which position Dylan Larkin plays or some other goofy misunderstanding. However, it seems that if you ask your average Red Wings fan whether these two should ever be on a line together, the answer is apparently that they mix like oil & water or like Maple Leafs fans & deodorant.

Winging It In Motown @wingingitmotown Follow

Poll time: Do Tomas Tatar and Gustav Nyquist work together as linemates?

11:59 PM - 23 Aug 2016

46%	Yes
54%	No

804 votes - Final results

TRENDING

Red Wings sign Christoffer Ehn to entry-level contract

Image 3.26: Presumption of shared knowledge 1

coaching, management, and even training, equipment, and other staff, prospects, and franchise history. Knowledge of these aspects on the part of the reader is often assumed by a blogger when he or she writes a post, and it is often assumed by commenters as well, unless a reader explicitly expresses a lack of understanding.

OPINION

The NHL is going to implement a John Scott Rule for All-Star Game Voting, and that's ok

The magic happened despite the league, not because of it
 by J.J. from Kansas | @JJfromKansas | Nov 1, 2016, 2:00pm CDT

TWEET SHARE PIN REC





Photo by Sanford Myers/Getty Images

Last year's All-Star Game was one of the most-memorable in my life, in any sport. It also happened on a lark that got out of everybody's hands so fast that it snowballed into a setup that was either going to be magical or disastrous. The career goon John Scott won the joke vote, then won a chance to even play in the game, then won the people over, then won the MVP of the whole contest.

The entire time this was happening, the sponsors-first, casuals-second, die-hards last National Hockey league was scrambling as best they could to control, mitigate, and eliminate the situation their own fans had put them in. Meanwhile, the voters watching on both sides were busy flashing their fan-police badges at one another over various charges related to the oscillating degrees of sanctity held by the All-Star Game, the voting process, the anti-goon movement, and even the man who had been thrust into the center of the entire situation.

At the end of it all, we got a story worthy of a movie deal, literally.




KIA SPRING SAVINGS TIME

2016 "Highest Ranked Small SUV in Initial Quality" by J.D. Power

Save! VIEW OFFERS

TRENDING



Red Wings sign Christoffer Ehn to entry-level contract

Image 3.27: Presumption of shared knowledge 2

In one post, blogger JJ From Kansas discussed whether two wingers, Nyquist and Tatar, truly play poorly together. During this discussion, he makes reference

to another Red Wing, Dylan Larkin, suggesting that if fans are uncertain about whether Nyquist and Tatar are the Wings' two best wingers, it's "likely due to a confusion about which position Dylan Larkin plays or some other goofy misunderstanding." In order for a reader to understand this reference, they would have to know that Dylan Larkin plays the position of center, and thus cannot rank above Nyquist and Tatar in the position of winger, as these are generally considered mutually exclusive. Another post, also authored by JJ From Kansas, focuses entirely on the NHL's handling of low-level physical player John Scott winning a fan-voted

Ken Holland opened this Red Wings off-season by insinuating there would be **change**. Henrik Zetterberg soon **echoed** those statements. If your general manager and captain speak of change, you assume it's coming, but when one of the first signings of the off-season is extending Drew Miller, you take pause.

Holland is coming off a draft that has been hailed by most as a **success**, opening up cap space by trading Pavel Datsyuk's contract to Arizona without having to move any tangible assets.

Now armed with an extra \$7.5 million to play with, Holland declared to the hockey world he was entering the Steven Stamkos sweepstakes. Which is good, I suppose, because adding a generational talent in Stamkos would create some buzz around the organization that hasn't existed since Holland went all in on a one year contract for Marian Hossa some 8 years ago. It's making a serious attempt to add a perennial Rocket Richard Trophy contender--someone who you can bank on for 40 to 50 goals. At face value it's difficult to argue with.

Holland actually executing a frugal trade at the draft? That's some good--take that to the Coinstar machine and GET PAID--change.

Holland pushing all the chips in to go after a superstar? Um, seems like a change because it's been a **few years**, but actually no, not a philosophical change. This is actually a very Ken Holland/Mike Ilitch thing to do.

Re-signing Drew Miller? NOT CHANGE. So the opposite of change, being that Miller embodies exactly what this team needs to change, it's stale as the bread sitting on top of my refrigerator.

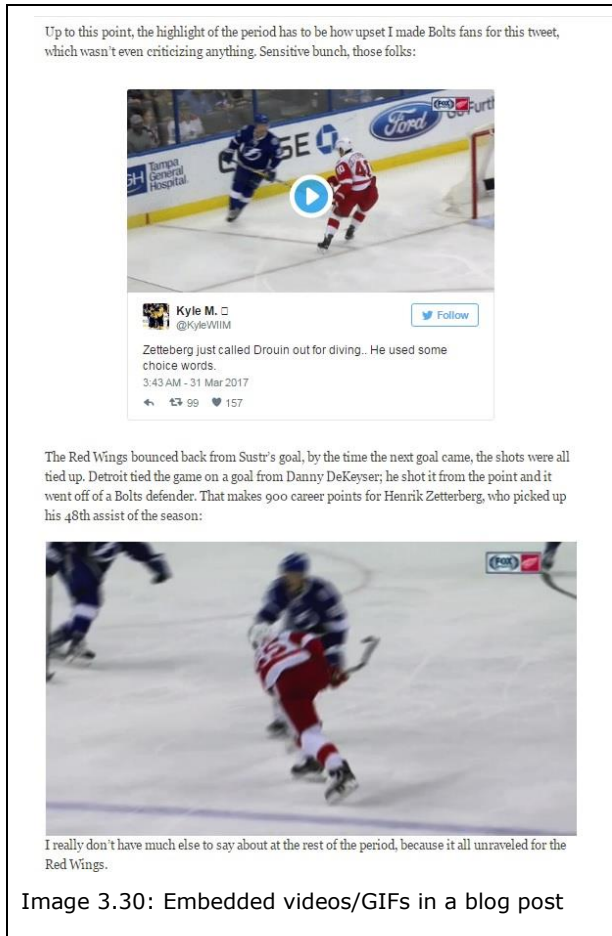
Holland has made a bed of his laurels and uses players like Miller and Luke Glendening as his pillows. It's clear, Holland and Jeff Blashill both believe in the 1998 ideal of needing a "checking line" to complete a roster. That you have to play defensively even when tied or behind. Holland and Blashill also (seem) to fall prey to the "eye test" which often leads you to believe a player like Miller puts forth more effort than say Gustav Nyquist, because instead of being in control, he's constantly chasing, constantly covering for little mistakes and often over matched. *WJIM's* Prashanth Iyer and *WJIM* editor JJFromKansas' thorough **breakdown** of penalty killing deployment has shown that Miller has **under performed** shorthanded--more so a "penalty killing specialist" is nearly impossible to define. Iyer also pointed out in a series of exasperated tweets that the Wings PK has slowed down through out the season **even in years** Miller has been healthy. Further damning, Miller's even-strength possession stats for his injury shortened 2015-16 season have shown him to be one of the worst forwards in hockey (**38%CF**).

If analytical breakdowns bore you, take a minute to look at this **cool website** that shows even his traditional stats have failed to hold weight the past couple seasons.

When I heard of "change", I was hoping a player like Miller would respectfully part ways with the Red Wings in favor of more youth, speed and skill in the bottom six and penalty kill. The Wings have more than enough young talent in their organization seemingly stuck to fill a void left by a player who measures out to be sub-replacement level. It would be a philosophical change away from the grinder, away from loyalty equaling merit.

With the unrestricted free-agent signing period beginning Friday, there is still much to be determined before reaching a final judgement on Holland's off-season. Having one of your first signings of the summer be Drew Miller was a static, if not regressive start.

Image 3.28: Presumption of shared knowledge 3



heavily used. Images, GIFs, and video are also commonly embedded in posts, and in fact almost every post on the main blog includes a title image- an image at the top of the page which is in some way relevant to and often helps illustrate the primary topic of the post. These multimodal forms are typically used as support for the written content of the blog post, but they are also occasionally used to carry their own communicative content, as a stand-alone communicative feature. This is especially common in the comments

section but also occurs in the main body of a blog post at times. In these situations, there is not necessarily any written context directly connected to the multimodal feature, as the feature conveys all of the meaning on its own.

The medium of the blog is a permanent one, as each blog exists on its own web page which can be permalined- a direct link to that specific blog post's page exists and can be shared. These pages are housed on a server, and while in theory they thus only exist as long as that server remains functional and continues to house them, there exist several online utilities which archive internet pages frequently, allowing accessibility to web pages that are no longer available on their original server. Thus, WiiM's posts are saved and accessible via these internet archives even

if the server ceases to house them or function at all. WiiM, like all blogs, is considered an electronic form of writing, making it easy to edit, even after posting, and easy to disseminate.

5.4 Production Circumstances

Blog posts are generally not composed live with reader access available as each letter is typed, and this is also the case with WiiM posts. They are composed privately and are generally planned pieces, with the author taking his or her time composing them and editing when necessary. The text only becomes publicly accessible once it is published as a more or less finished product on the blog. Bloggers can, however, edit a completed and posted blog at any time if they so desire. *The Managers* can also access and edit the blog posts of bloggers lower on the hierarchy, although that practice is rare in reality, as they have administrative access to all posts across the blog, including the main blog, Fanposts, Fanshots, and comments.

The posts themselves are asynchronous, as are many of the comments. As previously discussed, the comments sections of the game threads are designed to function as a more synchronous, chat-style form of communication, as opposed to typical blog comment sections which generally function in an asynchronous way. On popular posts, however, comments can still be posted and reacted to so rapidly that they are essentially functioning in this synchronous chat-style manner as well. This is likely due to WiiM's quite large and highly interactive audience.

The fact that the text from WiiM's posts is primarily composed privately and without the demand of time constraints placed on synchronous communicative methods, especially speech, is very important to a register analysis of the text and

even more so to further endeavors with the text, such as authorship studies. Bloggers have time to prepare the post in its entirety, to edit and fashion it to their liking, even if that means multiple edits over an extended period of time. They could spend hours or even days composing a post. The amount of time they have to create the text as well as the ability to edit and the knowledge that they can edit if they need to may have a notable impact on their linguistic choices. The lag time between production and consumption of the text permits the author to edit out, for example, mistakes that may be common in their writing and thus idiolectal in nature, either manually or with the assistance of software. This process could potentially remove indicators of the author's idiosyncratic tendencies, their common habits when writing or otherwise composing language. Synchronous communication largely removes the producer's ability to do this due to the demands of composing language quickly, listening to or reading a response in a timely fashion, and then being prepared to respond yourself in a timely fashion as well. Synchronous communication carries the expectation of rapid production and rapid reply and simply doesn't leave the language producer much time to thoroughly consider or edit their language.

5.5 Setting

The setting of the communication is also a vital aspect of a register analysis. Typically, in the case of many blogs and in the specific case of WiiM, the time and location of the production of the posts is not shared by the blogger and the reader. The blogger may or may not be entirely alone while composing, but if others are present they are unlikely to be members of the WiiM audience, and if they are, they are still not likely consuming the post as it is being produced. As previously stated, the posts are composed in their entirety in a private setting by the blogger, and only after they are completed and edited to the blogger's liking are they published on the


blog and thus disseminated for public consumption. There is a lapse of time between production and consumption and the time of production is thus also not shared between the blogger and reader. Comment sections in game threads are the exception to this. While location of production is generally not shared amongst participants, time of production arguably is, as participants are generally all reading comments as they are sent and replying in real time, discussing live events as they occur and at times relying on that context for comprehension of the discussion. This provides a strong argument that the time of the communication is shared by participants on both sides in the same way that it is shared by people in live chats or active messenger conversations.

The bloggers compose their blogs in a private place of communication, though the place of the communication becomes public once the post has been published to the blog. Once it is published, the post is readily accessible to anyone with an internet connection for as long as it remains on the internet, either on the original page or in the form of an internet archive. The reader's place of communication may be either private or public, depending on their circumstances at the moment. The specific setting in which the communication itself exists is the blog page, the digital environment itself. The physical locations in which the blogger can write posts and the reader can consume them are innumerable. The posts can be both written from and read from just about anywhere, though all are ultimately produced and consumed through the digital environment that is the blog page. The blog posts are written in a contemporary time period, as opposed to a historical one. The posts used to create the corpus for the examination of linguistic features in this analysis were created within two years of this writing, from October 2015 through October 2016. The very first blog post posted to WiiM is dated August 16th, 2007, about one

Report: Red Wings interested in moving Axel Holmstrom to North America

Sounds like the Swedish standout center could be on the move.
 by Kyle WiM | @KyleWiM | Mar 30, 2017, 10:56am CDT

TWEET SHARE PIN REC

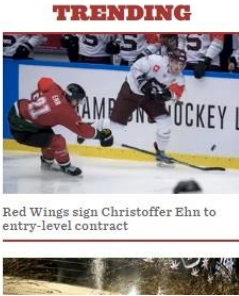



DETROIT, Mich — Reports from overseas indicate that Red Wings prospect Axel Holmstrom could be on his way over to North America now that Skellefteå's season is over after being eliminated by Christoffer Ehn and Frölunda HC.

"I have received signals that they want of me, leaning in that direction," reports the **Sports Bladet**, on Wednesday. Holmstrom's contract situation with Skellefteå remains to be seen, but the Red Wings previously expressed interest and patience with his development in Sweden. From what I've heard from a source in the NHL is that they're ready to have Holmstrom begin his development in North America now that the team is poised to take a new direction; a la rebuild and reload.

Holmstrom spent much of his season injured, but continued to be a playoff producer for his team, posting six points (3-3-6) in six games before being eliminated. Holmstrom is your prototypical two-way center who plays with an edge to his game, a style of play that typically translates well in North American hockey. The article above gives the notion that Detroit has interest in bringing him directly to the NHL club, but that would go against how they've operated over the past several seasons. My bet is that they will try to load up Grand Rapids with as many of their talented prospects so they can poise for a deep playoff run.

Image 3.31: Link to external content- informative



Red Wings sign Christoffer Ehn to entry-level contract

decade ago.

5.6
Communicative Purpose

Biber and Conrad's register analysis requires describing communicative purpose in general and specific terms. WiiM's blog posts cover multiple general purposes,

depending on the type of post. Posts in the category of Quick Hits are designed to curate internet content. The post body consists entirely of links to other content on the internet, which is generally relevant to, either the Red Wings or hockey in a broader context. The purpose of these posts is to help the reader stay informed regarding discussions and news beyond the blog, which will help give them the necessary specialist shared knowledge and context to navigate the contents of the posts successfully, as some may make reference to circumstances found in those links. In these posts, unlike almost any other post on the main blog, there is no editorializing on the part of the blogger doing the posting. If a blogger wishes to

Ken Holland Remains Defiant As Rebuild Talk Intensifies 80

Detroit's general manager pumps the brakes on Red Wings' rebuild.

by jmcurren28 | Mar 31, 2017, 3:00pm CDT

TWEET SHARE PIN REC



Photo by Bruce Bennett/Getty Images

Earlier this week I wrote an article covering comments made by the Detroit Red Wings' senior vice-president, Jimmy Devellano, following Detroit's elimination from the 2017 NHL playoffs. Devellano stated the "The rebuild is on" in Detroit, officially appearing to admit that the powers that be in Hockeytown may have seen the errors of their recent ways.

During an interview with Sportsnet on Thursday, Ken Holland seemed to pump the brakes on the whole 'rebuild' talk:

"I think we want to be competitive," Red Wings GM Ken Holland said on Prime Time Sports Thursday evening. "I'm a general manager and as long as I'm a general manager I want us to be the very best we can be. I don't believe in tear downs and massive rebuilds because I don't believe you can just guarantee the end result is going to turn out to be Stanley Cups and dynasties. You could go in the wilderness."

Yes Mr. Holland, you are indeed the general manager of a professional sports team. As a general manager, it is both your job and duty to build a competitive and successful squad. But with only one more year remaining on your contract, you should not fear the uncharted wilderness.

Image 3.32: Link to external content- editorialized

A vertical advertisement for CenturyLink Prism TV. The background is a solid green color. At the top, the text "Get CenturyLink Prism TV today." is written in white. Below this, there is a small orange button with the text "LEARN MORE". At the bottom, there is a small logo for CenturyLink Prism and a disclaimer: "Service not available everywhere. Residential customers only. Restrictions apply."

TRENDING



Red Wings sign Christoffer Ehn to entry-level contract



actually discuss or comment on a link to content found elsewhere, they will create a full blog post with the link to that content and then write their own story on the topic. These posts may involve a summary of the linked content or the

blogger may editorialize on the topic by expressing their own opinion or stance. They may narrate or explain, without personal stance, the topic of the linked content as well. Other posts do not link to any external content and may serve a variety of purposes as well. After each game, a post is uploaded wherein the author has essentially live-blogged, in private, the happenings during the game. The post is broken down by periods and is uploaded sometime after the game or early the next day. Sometimes included in these posts are embedded GIFs or video clips showing specific events mentioned in the text. These posts are designed to report significant


GETTING TO KNOW THE NHL RULEBOOK

Getting to Know the NHL Rulebook: Intent to Blow

Yes, in fact, there ARE legitimate cases for the infamous "Intent to Blow" rule.

by [wsgt2hhdms](#) | Jul 22, 2014, 8:00am CDT

Tweet Share Pin Rec



Christian Petersen

NHL Official Rules 2013-14 (PDF)


Section 5 - Officials

Intent to Blow the Whistle

This isn't actually a separate section, but given how big an issue it is, it warrants its own heading (and its own post). First, let's consider the offending paragraph in Rule 31 - Referees, section 2 - Disputes:

As there is a human factor involved in blowing the whistle to stop play, the Referee may deem the play to be stopped slightly prior to the whistle actually being blown. The fact that the puck may come loose or cross the goal line prior to the sound of the whistle has no bearing if the Referee has ruled that the play had been stopped prior to this happening.

TRENDING



Red Wings sign Christoffer Ehn to entry-level contract

Image 3.33: Informative- Getting to Know 1

events that happen during the game for readers who may have been unable to watch, as well as to inform, report, and describe. The blogger may include some minor editorializing amongst the reporting of events, and this personal stance taking can also be for the purpose of persuading the reader toward the opinion of the blogger. WiiM also has three post series, *Getting to Know the*

CBA, Getting to Know the NHL Rulebook, and Getting to Know General Advanced Stats. The posts in these series break down complex topics and present them in less technical, more layman-friendly ways to make them more approachable to average fans. Some include video clips to help illustrate more complicated topics as well. These posts are overall

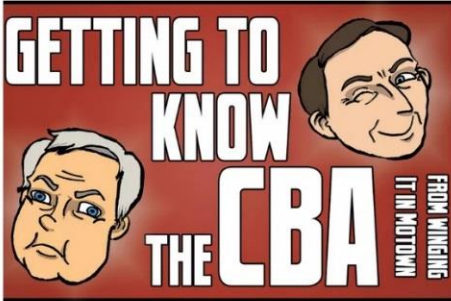
GETTING TO KNOW THE CBA

Getting to Know the CBA - Episode 4, Part 1: Standard Player Contracts

Thanks to the incredible Stevie Roxelle for the new banner logo for GtKtCBA

by [J.J. from Kansas](#) | @JJfromKansas | Jun 28, 2013, 6:00am CDT

Tweet Share Pin Rec



Stevie Roxelle - @stevieroxelle - Biscuit Fox


We've made it through 10 of 50 articles to the NHL's new CBA and have covered how the draft works, what makes a free agent restricted or unrestricted, and why they have a CBA in the first place. Today, we'll look at one of the larger articles, this time dealing with how teams and players agree on the contracts which define how every team is put together.

You can find the entire CBA here (PDF)

Article 11 - Rules and Procedures Governing Standard Player's Contract

If you remember back to the end-game of the lockout, when the players voted in favor of allowing the NHLPA to disclaim interest, the NHL followed up by filing a lawsuit in federal court. This lawsuit asked (among tons of other things) the court to judge that, since all existing player contracts were signed under a collective bargaining agreement, that the lack of a CBA in place would make all of those contracts void.

TRENDING



Red Wings sign Christoffer Ehn to entry-level contract

Image 3.34: Informative- Getting to Know 2

Methods:

To ensure that every player had at least one season of data, players with at least 500 minutes at 4v5 between the 2007-2016 seasons were included. The resulting dataset included 187 forwards and 191 defensemen. From [Corsica.hockey](#), data on each player's 4v5 time on ice, defensive zone faceoff starts (DZS), on-the-fly (OTF) starts, Corsi against per 60 (CA60), Fenwick against per 60 (FA60), shots on goal against per 60 (SA60), goals against per 60 (GA60), and expected goals against per 60 (xGA60) were obtained. For more details on the expected goals model, please see Emmanuel Perry's complete write-up [here](#). CA60 was selected as the primary comparator given that the ice time samples in this study are relatively small despite incorporating all available data.

Each player's 4v5 OTF shift starts per 60 minutes (OTF60) were plotted against his CA60, FA60, SA60, xGA60, and GA60. A linear regression line of best fit was plotted for each comparison. The r^2 were calculated for each comparison.

Results:

Figure 1 demonstrates the relationship between OTF60 and CA60 for forwards.

Figure 1. OTF60 vs. CA60 for forwards (>500 mins), 2007-2016

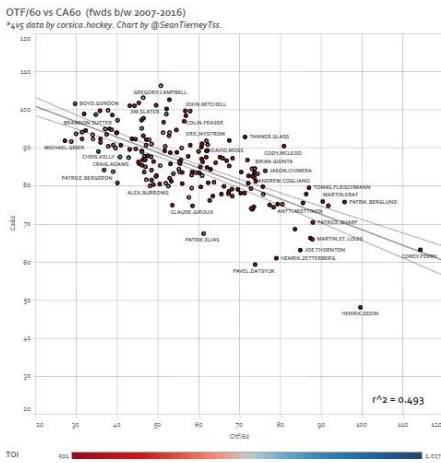


Image 3.35: Analytics/statistics post

readers in developing knowledge of the franchise, both present and historical, as well as a deeper understanding of the league and the sport. Bloggers may also find themselves learning new information, when readers supply information in comments or Fanposts that is new to the bloggers. As a platform for knowledge exchange, the blog provides a place for participants from both sides to learn and to contribute knowledge.

free of personal opinion and are designed specifically to describe and explain.

These general purposes lead into more specific communicative purposes driving the blog posts. WiiM posts often collect information from other sources and summarize and synthesize it in one place. For both bloggers and readers, the blog functions as both a learning platform and a teaching platform. The bloggers provide information and discussion that assist

The Red Wings host the Devils in the final game both of these teams will play this season. With both teams out of the playoffs, what do you play for now? Well, as Zetterberg said, they play for the Winged Wheel. And the Devils? Well, the Devils are not too far removed from a 6 game losing streak, definitely on a downward spiral as the season ends.

Tonight isn't really about the game; it's about celebrating the end of an era and all the people to whom the Joe was home. There will be laughs, there will be tears, but above all else there should be hope. This is certainly the end of an era, but a new chapter of Red Wings history begins next season. The LCA may be shiny and new, and there are some shiny and new Red Wings players to go along with it, but at least for a while guys like Zetterberg will still be driving the bus, and the rafters will have plenty of decoration to honor the past.

What amazing moments of Red Wings history will take place at LCA? What players will become legends? What new banners will be raised? I look forward to seeing how things develop, but before we move on we have to say goodbye to the past.

Draft position be damned, let's win tonight and send the Joe out in style.

Reports say that Andreas Athanasiou may play tonight, but we'll have to wait until later to know.

LET'S GO RED WINGS!

Red Wings

Forwards

Tomas Tatar	Henrik Zetterberg	Gustav Nyquist
Justin Abdelkader	Dylan Larkin	Matt Lonto
Darren Helm	Frans Nielsen	Riley Sheahan
Drew Miller	Tomas Nosek	Ben Street

Defensemen

Dan DeKeyser	Nick Jensen
Niklas Kronwall	Mike Green
Xavier Ouellet	Robbie Russo

Goaltenders

Jimmy Howard
Petr Mrazek

Scratches: None

Injuries: Andreas Athanasiou, Anthony Mantha, Luke Glendening, Ryan Sproul, Jonathan Ericsson, Johan Franzen (LTIR), Joe Vitale (LTIR)

Image 3.36: Pregame post

OPINION

Make Jimmy Howard Trade Bait

144

It's important to not only trade Howard this offseason, but to do so as soon as possible.

by MikeyLikeyHockey | Apr 8, 2017, 10:09am CDT

TWITTER SHARE PINTEREST RECOMMEND







Photo by Gregory Snamus/Getty Images

With the Red Wings officially eliminated from the post-season, the onus is now on management to make this team's future as bright as possible. That starts with Ken Holland wheeling and dealing beginning April 10th, moving anything with short term value for assets with long term value. One of those assets moved needs to be Jimmy Howard.

TRENDING



Red Wings sign Christoffer Ehn to entry-level contract

Why Move Howard?

With no certainty in the Expansion Draft, the Wings could possibly lose a valuable player for nothing if Mrazek is protected (and despite some abysmal stats this year, I still think he should be). Howard will be ready to retire, or at least will be declining, when this team is ready to compete again in what will probably be several seasons - and that's on the best timeline. We need prospect and pick assets, so let's move player assets.

Jimmy Howard (\$5.29M/yr cap hit) is possibly the most valuable asset left on the team not named Mantha, Larkin, or Athanasiou. Over his career, Howard has put up a 2.44 GAA and 0.915 SV%. But, this season, he's sporting a 2.16 GAA and 0.927 SV% - those numbers are good for 4th and 2nd best, respectively, among all goalies who have played at least 20 games this season. This should be very valuable to a team that sees itself as only a goaltender away from contention.

One blogger of *The Editors* frequently uses blog posts to perform and discuss advanced statistical examinations of the players and the team overall. Pregame posts are posted before each game which include preparatory information on the up-to-date state of the Red Wings team at that time, including

Why Does This Need To Happen Soon?

The Expansion Draft is going to throw a wrench in at least a couple goalie situations. While I'm not convinced that Jimmy Howard will be our player selected to go to the Golden Knights, the possibility certainly can't be ruled out. As an asset with trade value, we're realistically looking at possibly losing him for nothing or paying Vegas with draft picks to not select him. Why do that when we can send him to a team that doesn't have a goalie worth protecting?

And the sooner we can ship him out, the better. As soon as they're eliminated from the playoffs, teams like Washington and Pittsburgh will be looking to deal goalies like Philipp Grubauer and Marc-Andre Fleury so that they don't lose net-minder assets for nothing in the Expansion Draft. With the Wings and Stars both on the outside looking in, the two teams could make a trade as soon as April 10th. If I had Ken Holland's ear, I'd have him tell Jim Nill that it's better to deal with the Wings immediately at the end of the regular season than to get involved in a bidding war that just makes some of the NHL's best teams even better long-term.

Ken Holland has done a wonderful job of selling snake-oil to Red Wings fans the past few years. This off-season, it's time for him to turn those skills on his own players to get some clauses waived and veterans moved. Start the sales pitch on Howard, Kenny...

Image 3.37-3.39: Post expressing blogger opinion

line up and goalie choices, as well as information regarding the opposing team. These posts help prepare the reader for that day's game.

Similarly, post-game posts are uploaded after every game to summarize the game and present the final outcome for readers. The game thread posts provide a place for fans- including both bloggers and readers- to come together during games to discuss live action in a chat-style manner. Comment sections on other posts also provide readers a place to voice their opinions on the topics discussed and those parallel. Fanposts offer them a place to expand those discussions. The bloggers also use the blog as a platform for presenting their own ideas and opinions about the team and the sport, and their social positioning within the blog places them as experts on these topics, leading readers to highly value their arguments.

A recent rift has arisen between a number of bloggers in the Red Wings online community, led by the WiiM bloggers, and professional beat writers covering or working for the team. The bloggers, supported by many readers, suggest that the beat writers are not critical enough of or honest enough about the team and that they do not demand answers to the questions that their readers want asked of team



Image 3.40: Post critical of mainstream beat writer 1

officials and players. The bloggers have gone so far as to post entire blog posts discussing the issues they have with these mainstream media members, and the situation has grown

Free Press writer Helene St. James has an idea. While she initially praised Mrazek for his improved play as of late, she also suggests that this may benefit the team more than with an improved position in the standings. Here's my favorite tidbit:

If Mrazek, who turns 25 in February, has a hold on the starting job by then, good for him and good for the Wings. The thinking within the club is that Coreau is ready for the NHL full-time, and he may well end up being the goalie the Wings protect in the expansion draft this summer. If Mrazek gets on a roll – remember last year around this time

That sound you hear is two things:

- Cheers from the ever-present crowd in Detroit whose favorite athlete is “whoever is backup goalie for the Red Wings”
- A large crowd charging Joe Louis Arena demanding to “speak to Ken Holland”

Fear not, my friends, because if there is one thing I have learned in my years as a Red Wings fan, it is to take articles from this specific Detroit digger with a fistful of salt.

What I see in that little nugget of information are two separate ideas: one, that Coreau is ready for a full-time NHL position, and two, that he may be protected this summer. It's made clear that the Wings' management believes in the former. I'm not at all convinced for the latter. To me, that sounds like a lot of speculation we've seen from Ms. St. James in the past that, to put it bluntly, never made it to fruition.

Look at the Red Wings' roster and think back over the last couple of seasons. I still see **Gustav Nyquist** and **Tomas Tatar**, and nobody named **Dion Phaneuf** on that list. Yet multiple articles by the so-assumed mouthpieces for the organization were written of the seemingly inevitable trades of and for those mentioned players.

Valteri Filppula wanted oodles of cash and a role as a first line center, right?

Image 3.41: Post critical of mainstream beat writer 2

If it is the latter, don't be surprised to see Mrazek in play. It's what makes sense, as Coreau has impressed better than his numbers (3.04 GAA, .901 save percentage) and Howard (1.96 GAA, .934 save percentage) has played the best hockey of his career.

“The latter” here being the Red Wings in a position to sell once the trade deadline comes around on March 1.

Pay attention to the language used here: “it's what makes sense.” Sense to whom? Apparently sense to the writer, but not necessarily to the Wings' brass. St. James goes on to throw cap hits and ages into the question, citing that Howard's contract will be more difficult to move given those factors and despite his excellent play.

While that is true, I don't see the team making a trade involving Mrazek at the deadline. They just committed \$8M to him over two years, pegging him as their number one goalie. Hinging their future on one down season for their main man, the unforeseen resurgence of a 32-year-old before yet another injury, and ten games by a still-learning rookie isn't the Detroit style for goalies.

As it stands, we're long past the point we thought we would be at with both Howard and Mrazek still on the team. The Red Wings simply don't make quick decisions on their goaltenders.

So considering all of that, what we've seen from HSN in the past regarding asset/contract management and what actually happens, I think it's as likely as **Jonathan Ericsson** winning the Norris trophy this season.

Image 3.42: Post critical of mainstream beat writer 3

vitriolic, with beat writers muting and blocking bloggers on social media and refusing to interact with them. Blogger redwinger43 wrote an entire blog post blasting beat writer Helene St. James for her take on a recent issue with a goaltender as well as multiple previous situations.

Redwinger43 declared that she has learned to “take articles from this specific Detroit digger with a fistful of salt.” She then proceeded to bring up previous instances in which St. James wrote articles that, in retrospect, were shown to contain likely inaccurate

information or ill-founded opinion, such as when she claimed that player Valtteri Filppula wanted “oodles of cash and a role as a first line center” to re-sign with the team only to see him sign with a different team at a reasonable cost and in a second-line center position. In recounting these situations, redwinger43 appears to be mounting evidence to undermine St. James’ expertise regarding the team.

WiiM also offered a platform for a very prestigious former Wings blogger, Michael Petrella, to lambast the mainstream Detroit media and particularly the beat writers. Petrella claimed that “no one is willing to rock the boat or burn whatever bridges they perceive they have,” that “they refuse to criticize,” and that “no one has the guts to question” the team or the answers they give. The amount of support the bloggers have received on this issue illustrates the regard with which much of the online Red Wings fan community holds them. This suggests their status as trusted

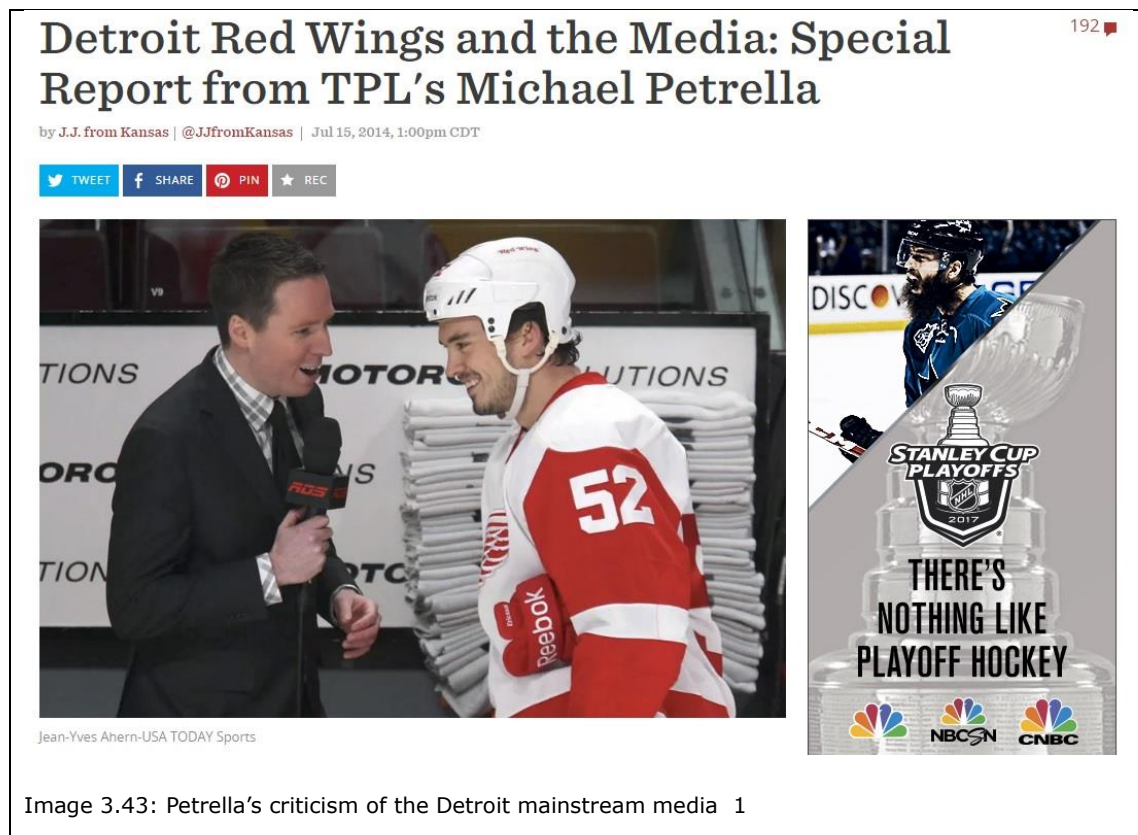


Image 3.43: Petrella’s criticism of the Detroit mainstream media 1

Once upon a time, I had a good relationship with the Detroit Red Wings. I had a two hour breakfast with Ken Holland, where he was very open and never once asked me NOT to print something. I had an open invitation to sit in the press box if ever I was around for a game (I no longer live in Michigan, and haven't since I graduated high school). I was briefly credentialed by the NHL, and had the opportunity to interview Red Wings draft picks in Los Angeles at the 2010 Draft. I was even approached by the Red Wings to publish some information about a move that didn't get a lot of positive publicity, though I was asked not to say where the information came from.

But all of that stopped. I can't explain why -- or what happened. But it was abrupt. I'm pretty sure that it's based in The Production Line's increasingly critical tone about Red Wings management and on-ice product. All I know is that one day e-mails stopped being responded to, the press releases stopped coming into my inbox, and things I'd discussed doing with the Red Wings were being done by other people.

When news breaks -- like the Dan Cleary signing, for example -- it's announced practically verbatim by several writers. Every article is nearly identical, but not identical enough to just be the press release, and is published at precisely the same time. It's the hockey equivalent of state-run media. It seems as though the only information that's released by the allegedly-independent media is the information that the organization wants to be released. Is it that anyone that isn't willing to toe that line -- bloggers, included -- are excluded from access to the team, its players, or members of the front office?

No one seems to have a problem with that. No one has the guts to question it. No one is willing to rock the boat or burn whatever bridges they perceive they may have.

So I figured... I've already been blackballed, what do I care? The mainstream media -- or diggers, as they're passionately known by the Red Wings community -- doesn't care, presumably because any deviation from the company line will cost them their access. So, instead of doing what they've committed to doing -- reporting, asking tough questions, and making good on their journalism degrees -- they do nothing. They refuse to criticize, and happily post the exact same thing that all of the other writers in the area do. Pretty groundbreaking stuff.

I reached out to quite a few people to talk about it. Former members of the Detroit Red Wings media departments, current and former Detroit Red Wings beat writers, as well as national hockey media. Will there be any effect? Probably not. This seems like it's become accepted practice, but I couldn't sit by any longer, only reading what the North Korea of NHL teams wants its citizens to believe. But it's worth a shot -- and if I learned anything from my time running TPL, it was that no one answers if you don't ask.

And, usually, the only question that needs to be asked -- and no one ever seems to ask -- is "why?"

Image

3.44-3.46: Petrella's criticism of the Detroit mainstream media 2

I asked the above-quoted former DRW media department employee, and everyone else, if the Red Wings threaten -- explicitly or implicitly -- to revoke their access to the team, its players or staff if they refuse to push the party line... and he makes it sound like the writers are more to blame than the team for pushing narratives:

The verbatim articles are a product of the writers, not the team. They're all chatty, friendly, with few rivalries amongst them. They are very much 'Super Friends' on the beat and are given limited crumbs in media availability, or on conference calls, etc. Unlike many [Original Six] markets, Detroit's media doesn't feast off of rumors, off-ice antics, or running players out of town. It is very much old school... this is the story... these are the quotes... this is the narrative I've weaved around it.

You could look at markets like Montreal and Boston, where the media will literally carve up and eat the players alive (Tyler Seguin, Tim Thomas) and say that what happens in Detroit is a good thing (or a bad thing), that's the way it is.

He even touches on some experiences with the team that I can absolutely confirm -- like the aforementioned Ken Holland conversation. That's evident when he says that "Babs will say no to a lot of things... Kenny won't turn down a reporter's phone call... and it's the only place where the 'Super Friends' do well."

Image 3.47: Petrella's criticism of the Detroit mainstream media 3

sources to be quite high. It also suggests that these fans feel the bloggers are providing a service for them that is going unfulfilled from the mainstream beat writers. This is a status level not often reached by internet bloggers and shows that another communicative purpose of WiiM is filling a gap in providing information and discussion to fans that the mainstream media may be leaving. This purpose elevates the status of the bloggers and very possibly changes the way they communicate via their blog posts and perhaps even what they consider the purpose of those posts to be.

Factuality is another dimension of communicative purpose which must be examined. As various posts have different purposes, so, too, do they meet varying levels of factuality. The series posts which exposit the CBA, the NHL rulebook, and

advanced stats are factual in nature. Opinion is a common level for WiiM posts to land on the factuality spectrum, with bloggers often writing pieces declaring their personal stance on topics such as the movement and development of specific

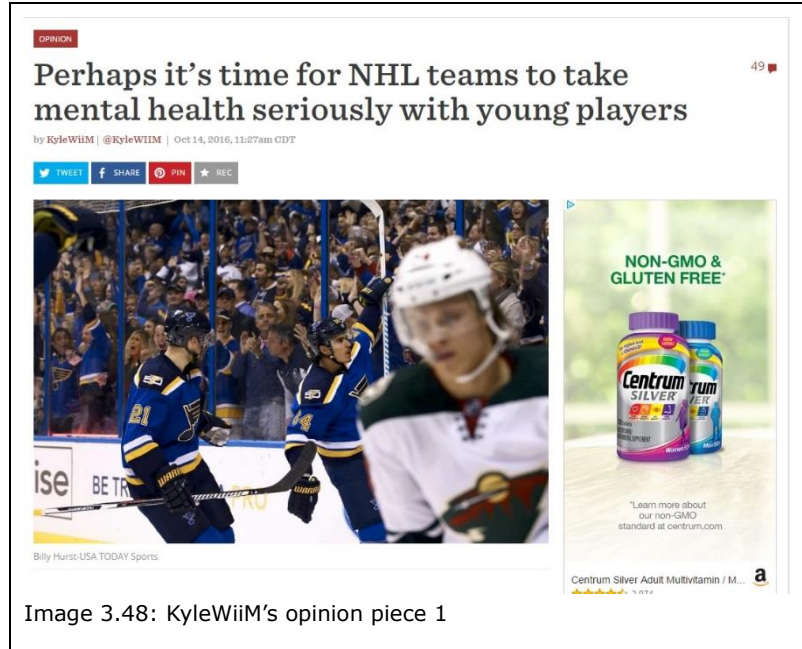


Image 3.48: KyleWiiM's opinion piece 1

prospects, trades, free agency signings, line up decisions, and player performances.

Blogger KyleWiiM even wrote a full post proffering his opinion that NHL teams should be putting more effort toward

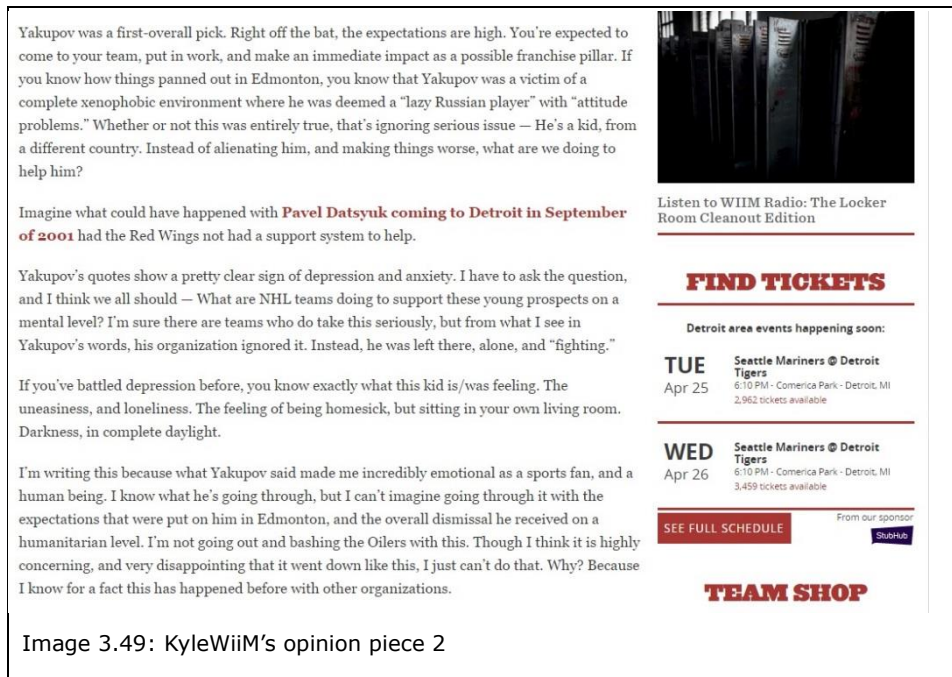


Image 3.49: KyleWiiM's opinion piece 2

supporting the mental health of young players, demanding to know "what are NHL teams doing to support these young prospects on a mental level?"

Predicting the Expansion Draft: Detroit

Ladies and Gentlemen, the moment you've been waiting for...

by PeterWiiM and MikeyLikeyHockey | Jan 19, 2017, 7:00am CST

TWEET SHARE PIN REC



Photo by Gregory Shamus/Getty Images

Image 3.50: Prediction post 1

Forwards

We predict that Detroit will choose the 7 Forward, 3 Defensemen, 1 Goalie option. They don't have four defensemen that are worth not being able to protect the three additional forwards.

Frans Nielsen must be protected because he has a No Move Clause. Even though his cap hit is high relative to his current level of performance, there is no way that **Henrik Zetterberg** is not protected, barring some unforeseen change. Regardless of any other arguments, if Zetterberg retires, even if he does so as a member of another team, before his contract expires, the Red Wings will get hit with a large **cap recapture penalty**.

At this point, let's address the one question everyone opened this article to read: Is either (or both) of **Andreas Athanasiou** or **Anthony Mantha** going to be exposed, and if so, what's the best bulk-pitchfork supplier?

Like many, if not all of you, Mike and I are of the opinion that these two are no-brainer protects. So, we are going to protect them, but we will later discuss possible reasons they may not be.

Barring any trades, **Tomas Tatar** and **Gustav Nyquist** should be the next slots for protection. Both should bring back a high enough return that it makes no sense to lose them for free, even if the plan was to trade one or both.

Our last protect is **Justin Abdelkader**. The team extended him through 2023, so we don't see the team exposing him to the expansion draft. This is similar to the DeKeyser scenario mentioned in the introduction. We would not protect him, but the team will.

Image 3.51: Prediction post 2

MikeyLikeyHockey co-authored even predicts which players the team will choose to protect in the upcoming expansion draft, using wording like "We predict that Detroit

Speculative posts are also common, such as predictions posts written by each blogger before the season starts, predicting how various aspects, such as wins and player performance, will play out over the course of the season, as well as posts predicting the progress and outcome of all teams in the playoffs. A blog post PeterWiiM and

EVENTS

WIIM's 2016-17 BOLD Predictions Volume III: The Veterans 19

by Peter WIIM | @jimmywellington | Oct 7, 2016, 12:30pm CDT

TWEET SHARE PIN REC



Photo by Taylor Weidman/Getty Images

It's that time of year again! Time for our writers to make some bold predictions about the upcoming hockey season. If you haven't experienced this series before, our writers break into groups of three or four, and each make five BOLD predictions for the 2016-17 season. We then respond to our other group members' predictions.



TRENDING

Image 3.52: Prediction post 3

JJ:

- Alexey Marchenko will lead the Red Wings defense in goals this season.**

Kyle: Gets put on waivers anyway

Graham: Red Wings as a team set a record for fewest goals as a defense corps with 3

Jeff: Only because Brendan Smith got traded.
- Shea Weber gets more Norris votes than PK Subban.**

Kyle: Habs MSM and fanbase will go nuts and think it was a good trade.

Graham: Both get fewer than Danny DeKeyser.

Jeff: This isn't really that bold of a prediction.
- Darren Helm scores at least five shorthanded goals.**

Kyle: You mean throughout the rest of his new contract, right?


Graham: You're like Charlie Brown and the football with this one.

Jeff: All of them on breakaways
- Jimmy Howard finishes the season with more shutouts than Petr Mrazek.**

Kyle: One is better than none, I guess.

Graham: For Edmonton.

Jeff: More losses too.



Expansion Draft Prediction Update - Part 1

Listen to WIIM Radio: The Locker Room Cleanout Edition

Vili Saarijarvi Nominated For OHL Defenceman Of The Year

Image 3.53: Prediction post 4

will choose the 7 Forward, 3 Defensemen, 1 Goalie option," "Barring any trades, Tomas Tatar and Gustav Nyquist should be the next slots for protection," and "We would not protect him, but the team will."

Both of these types of posts can be seen to fall under both the opinion and the speculative categories of factuality. WiiM posts also land on both sides of the expression-of-stance spectrum. Some posts are driven by the expression of the blogger's stance, while

others show little to no overt expression of stance at all. This is one aspect that complicates the process of placing WiiM on a spectrum of register types and

comparing it with other forms of text, as this factor can cause posts to vary widely in terms of linguistic features.

5.7 Topic

Topic is a very relevant factor to blogs, and WiiM is a blog that is very topic-driven and specific. The general topic domain for this blog is sports. More specifically, the blog is hockey-centric, with heavy focus on the NHL and primarily one team, the Detroit Red Wings. The majority of posts focus on the Wings in some way, and some venture beyond the team but often stay within the realm of the NHL. If posts do cover non-NHL topics, they are virtually always still hockey-related, covering topics such as Olympic and international/IIHF (International Ice Hockey Federation) hockey, European hockey leagues, or junior hockey leagues. Even these discussions are often focused on relevancy to the Red Wings, such as current or former Wings players or prospects- young players whose player rights are owned by the Red Wings but who have not yet joined the NHL team- who are involved in those non-NHL teams. The blog almost never strays beyond the topic of hockey in a general sense. True off-topic posts, even those regarding other sports, are very rare on the main blog. Among Fanshots and Fanposts, off-topic posts are slightly more common, but still quite rare. WiiM is very much a topic-driven blog. The comments sections see the most off-topic discussion, but even this area stays largely focused on hockey. The name of the blog, *Winging It in Motown*, is derived from the name of the team and its location, and that is the primary focus of this blog.

Chapter 4

Biber and Conrad's Linguistic Features

1. Introduction

I began the examination of linguistic features by building a corpus using blog posts from Winging It in Motown. An examination of linguistic features requires a body of text to examine, and compiling posts from the blog into a corpus format provides for the necessary data set for that examination. First, I explore the methodological approach to creating that corpus.

2. Creating a corpus

The first step in performing the linguistic features portion of a register analysis is to either find or create a corpus of texts to act as data. For the purpose of this research project, I created a corpus from posts on the blog Winging It in Motown. I chose to utilize 13 months' worth of posts, which would provide a reasonably large corpus in terms of word count. Doing so would also allow for the differences in types and number of posts from month to month based on specific characteristics of the hockey season to be accounted for, such as a particularly quiet period in August when little happens across the NHL, a spike at Free Agency and Draft times, and a shift from Wings-focused game posts to other teams during playoffs once the Wings are no longer playing. Culling data from an entire year allows for all of these changes in posting patterns, habitual across years of posts to the blog, to be included. All posts by all authors between October 1, 2015 and October 31, 2016 were gathered and combined to develop the corpus for this project. The choice to include all types of posts, with the exception of game threads

which have virtually no post content, was driven by the desire to gain a representative sample of the language of the blog overall, not merely for one subregister of post. The posts were collected in Microsoft Word with all images and visuals from within the posts included, and then the text was copied into Microsoft Notepad. This process removed the graphics, important for examination of the structure of the blog posts but unnecessary for the corpus data itself, from the text and allowed for the text to be run through linguistic utilities that work with .txt files.

Certain components of the posts were removed for the purposes of building the corpus. The author bylines, which included author name, date, timestamp, and author twitter handle were deleted from the text used for compiling the corpus. These aspects were deemed structural but not part of the language of the posts themselves, as they appear on every post in precisely the same manner. Text from buttons at the end of posts, such as the comment button and the variety of share buttons, was also removed. Cardinal numbers not part of the post text itself, such as those indicating the number of shares or comments, were removed as well. Bylines for image credits were also deleted, as this is once again a structural component and not part of the text itself. The decision was made to leave text from embedded tweets in the data, as these are similar to quotations from others and are thus part of the posts themselves rather than a structural component of the blog. While this may need to be dealt with differently for an authorship analysis of the same text, for the purposes of corpus data this is the appropriate option. Image captions were also left in place, as they are written by the blog authors specifically for that blog post. Formatting on the blog posts left certain portions of the text missing the appropriate spaces, so these spaces were manually re-introduced. Beyond this, the text was left unedited. All instances of spelling errors, grammatical errors, and typos were left as-

is. Once this clean-up process was completed, the text file was run through several software platforms for analysis.

3. Word Count

Antconc's tokenizer function was used to collect a basic word count of the edited corpus file. While counts were also obtained via wordcounter.net, the counts were slightly different, and as Antconc will be relied upon for counts of word types, I have chosen to report its raw word count as well. For the 13 months of posts collected, Antconc found 813,435 tokens, or individual- but not unique- words. The total number of posts from that time period that went into the compilation of the corpus was 1,549. By dividing the number of posts into the number of tokens obtained, the average number of words per post can be determined: approximately 525.1. This proved to be an interesting result in and of itself, as I (Cox, 2014) previously found the average words per post, after performing a similar examination on one month's worth of posts from March 2014, to be approximately 847.7. This is a large difference, a little more than 1/3 less, and suggests that the length of posts on WiiM has been significantly reduced over the course of just a couple of years, a result which came as a surprise. This may be partially due to the loss of CSSI posts, statistically-driven posts compiled by one specific blogger using his own system which were often over 1000 words and accounted for the majority of posts longer than 1000 words in the 2014 analysis. That blogger no longer compiles these posts, which were previously written for each game, removing a large portion of the longer posts from 2014. The average for this corpus of 525.1 words is, however, still notably higher than the 210.4 average words per post found by Herring et al. (2005) in their genre-based examination of blogs.

4. Sentence count and length

While Antconc's total word count is the basis for calculations of word types presented later in this analysis, Antconc does not offer sentence-level analysis, including a basic sentence count. For this count, wordcounter.net was used. Wordcounter found the total number of sentences in the corpus to be 39,197, approximately 25.3 sentences per post, again much shorter than the average 42.6 found in the 2014 analysis. The total number of words per Antconc divided by this total number of sentences gives the average number of words per sentence as around 20.8. This number is, in fact, slightly higher than the number found in the 2014 analysis, at 19.9, and remains well higher than Herring et al.'s (2005) finding of 13.2 words per sentence. This suggests that WiiM still seems to have longer posts with longer sentences than typical blogs.

5. Parts of speech counts

The corpus was run through the Stanford Tagger to determine parts of speech, with tag information obtained from Santorini (1990). The tagged corpus was then run once again through Antconc to obtain counts for each tag, which were then analyzed and, when necessary, compiled to obtain parts of speech counts. Biber and Conrad (2009) have shown that distribution of parts of speech varies, often dramatically, across text types. This data can thus help define what constitutes a specific type of text. Examining the commonality of specific text types is crucial to the establishment of the text in WiiM as a register and also aids in situating it among other registers by offering this data for comparison.

5.1 Nouns and Verbs

The Stanford tagger found this corpus to contain 284,943 total words marked as some type of noun, excluding pronouns. That accounts for about 35% of words in the whole corpus at a rate of 350.3 per thousand words. 130,932 words were tagged as some type of verb, accounting for only about 16% of the words in the corpus. The Stanford Tagger separates nouns into four categories: singular/mass nouns, plural nouns, singular proper nouns, and plural proper nouns. Verb tags are broken down into base-form verbs, past tense verbs, gerund or present participle verbs, past participle verbs, non-third person singular present tense verbs, and third person singular present tense verbs.

Part of speech	WiiM raw	WiiM frequency	COCA raw	COCA frequency
Noun, singular/mass	117027	143.8/1000	86691341	162.4/1000
Noun, plural	34576	42.5/1000	30196644	56.6/1000
Proper noun, singular	122704	150.9/1000	25574115*	47.9/1000*
Proper noun, plural	10636	13.1/1000	*	*
Total	284943	350.3/1000	142797324	267.5/1000

Table 4.1 Nouns

*COCA data on proper nouns was not separated for plurality, so data presented includes all proper nouns, regardless of plurality

Part of speech	WiiM raw	WiiM frequency	COCA raw	COCA frequency
Verb, base form	35131	43.2/1000	N/a*	N/a*
Verb, past tense	24262	29.8/1000	19885884	37.3/1000
Verb, gerund/present part.	17934	22.1/1000	7920309	14.8/1000
Verb, past participle	12983	16/1000	11231382	21/1000
Verb, non-3 rd p. sing. present	16654	20.5/1000	N/a*	N/a*
Verb, 3 rd p. sing. present	23968	29.5/1000	12404115	23.2/1000
Total	130932	161/1000	83743473	156.9/1000

Table 4.2 Verbs

*COCA data did not differentiate between base forms and non-3rd person singular present verbs (e.g., "You go..." and "You want to go...") so this data was not included in this study.

A. Regular and proper nouns

Counts for nouns and verbs in this WiiM corpus correspond closely to those obtained in the 2014 analysis. These numbers show the noun count for this corpus

as being on the high side, accounting for more than one third of the total number of words in the corpus. Biber et al. (1999) found nouns to account for only about one quarter of words on average, which aligns with the noun count for the Corpus of Contemporary American English (COCA), where nouns accounted for approximately 26% of all words (Davies, personal correspondence, April 29, 2017). Interestingly, the use of proper nouns appears to be quite high, with the count for singular proper nouns eclipsing the count for singular regular nouns. Only after plural versions of each are added in does the number of regular nouns exceed the number of proper nouns. Plural proper nouns are significantly less common than singular proper nouns, occurring at less than 1/10th the rate.

The high rate of proper nouns is less surprising when considering both the topic and the purpose of WiiM's posts. The topic- speaking both broadly, hockey, and specifically, one hockey team- lends itself toward the heavy use of proper nouns, from the names of leagues, such as NHL, Liiga, or IIHF, all the way down to staff members at the rink, such as Al Sobotka, building operations manager and head octopi twirler. Players, writers, teams, officials, NHL management, and a great number of other subjects with proper names are regularly discussed. In terms of purpose, WiiM posts are designed primarily either to inform or report or to express opinion about the topic. Both of these purposes relate to heavier use of nouns in general and an expected more frequent reference to proper nouns given the high number of proper nouns the topic introduces. While reporting on the occurrences of a game or explaining an analytic examination, for instance, the names of numerous players and teams would likely need to be introduced. Furthermore, because many posts are team-focused rather than focusing on one specific player, names would likely need to be re-stated multiple times as the blogger jumps around from

discussion of one player to discussion of another and back again, as seen in the following excerpts from a WiiM game recap post from March 30, 2017:

Let me just say, the name Yanni Gourde is not real. That is not a real person. There will be no debate.

Anthony Mantha got in a fight with Danny DeKeyser's former linemate in the opening five minutes, so that's cool. Petr Mrazek was tested early on, but the Red Wings got the opening goal shortly after the fighting major on Mantha:

...

In the closing minutes of the first period, it was announced that Anthony Mantha would not return due to an upper-body injury. Fighting is stupid.

The Lightning tied the game up on a goal from J.T. Brown with less than a minute left. Darren Helm's turnover led to the goal, Brown sniped it top-shelf on Mrazek's glove side:

...

Yep, Mantha is out for the rest of the season:

...

The chippy play continued to open up the second period. Nothing too exciting, but Tampa did manage to pull ahead on a goal from Andrej Sustr after he went to the net and a pass went off of his skate.

...

The Red Wings bounced back from Sustr's goal, by the time the next goal came, the shots were all tied up. Detroit tied the game on a goal from Danny DeKeyser; he shot it from the point and it went off of a Bolts defender. That makes 900 career points for Henrik Zetterberg, who picked up his 48th assist of the season:

...

Danny DeKeyser scored again, except, he scored on his own net, which sums up this entire season perfectly. The Bolts managed to score again — Jonathan Drouin danced through the Red Wings defense on a power-play brought on by a crosscheck from Danny DeKeyser.

...

The Red Wings continued to play a very lackluster game and the Lightning danced around them without a problem. The guy that I made fun of at the beginning of this so-called recap scored to make it 5-2.

Yes, Yann Gourde scored. Life continues to be a questionable hellscape.

The Red Wings went back to the man advantage, which to all of our surprise, they did NOT squander. Mike Green picked up his 13th goal with this effort off his own rebound with help from Frans Nielsen

The heavy use of proper nouns is evident in these excerpts, which account for about half of the total post. There are 11 different players and two different teams mentioned. The discussion regularly switches back and forth between the two teams, and more than half of those 11 players are mentioned more than once in separate instances with other players mentioned in between or where the referent is far enough back that the individual needs to be named again to assure clarity. Yann Gourde is discussed in the first sentence but then not mentioned again until the second-to-last excerpt near the end of the entire post. Eight other players are mentioned in between. Danny DeKeyser is mentioned in three of the excerpts, each of which also mentions at least one other player, necessitating the re-stating of DeKeyser's name in some form to avoid ambiguous pronoun antecedents. The purpose of this post is to inform the reader of the events during the game in chronological order, and in order to accurately and effectively achieve that purpose, the blogger must use numerous proper nouns multiple times.

B. Verbs

The overall high rate of nouns relative to verbs is also unsurprising when considering the register and its most common purposes. Per Biber et al. (1999), verbs tend to occur at a higher rate in registers that focus heavily on interpersonal relations, such as conversation. WiiM's posts, with their more informative and explanatory purposes, have a low focus on interpersonal relations and thus are likely

to showcase lower rates of verb usage. However, WiiM also includes posts of opinion and personal stance, and opinion and personal stance discussions sometimes occur in posts that are primarily designed to inform. This may explain why, while the rate of verbs is notably lower than the rate of nouns, the relative proportion of verbs to nouns is higher than the proportion Biber et al. (1999) reported for news and academic prose. Their findings indicated that in conversation the proportion of nouns and verbs is about half and half, while in news and academic prose the proportion is closer to three or four to one in favor of nouns. Proportions from WiiM posts fall in between those found in these registers, a finding which is unsurprising when considering the hybrid purpose of the blog overall. Biber et al. (1999) also suggested that the proportion of nouns to verbs reflects “the density of information packaging” (pp. 66). This theory matches with the hybrid purposes of WiiM as well. The informational aspect of the posts likely increases the amount of information to be conveyed, but the relatively strong assumption of shared knowledge holds that amount lower than would likely be seen in e.g. a newspaper report, where there may be little or no assumption of shared knowledge.

While the rate of nouns is much higher than the rate of verbs, the verb occurrence rate is still higher than Biber et al.’s (1999) finding that verbs account for approximately 10% of words on average. In addition to the previously-discussed hybrid purpose of WiiM, which includes opinion pieces with significant expression of personal stance, this may also be largely attributed to the topic of the blog. As Biber et al. (1999) stated, “lexical verbs denote actions, processes, or states and serve to establish the relationship between the participants in an action, process, or state” (pp. 63). Many of the blog posts report on the events that occurred during the games. Reporting on sporting events would likely see a heavier use of verbs because

of the highly active nature of the events being reported, as can be seen in the following excerpt from a post-game recap post from October 13, 2016, where 27 of the 150 words (18%) were tagged as verbs:

Forty seconds in and Tatar had a slick scoring chance, but it bobbed just out of his reach to really get much on the shot. Smith looked to be taking some initiative, too, jumping in early on, exactly the way Blashill wanted his defensemen to do when he first got his promotion to the big club. In the early going Detroit carried play, although Tampa was certainly gunking things up in the neutral zone as much as possible with their big bodies. Fortunately the Wings got some breathing room when Glendening managed to draw a holding penalty against Coburn. There were so many questions circling the special teams, especially the power play this past summer, as well as uncertainty the Vanek signing, but look at that, it took all of 20 seconds for Vanek to redeem all of the fears surrounding the man-advantage as he cleaned up Zetterberg's sharp-angle shot.

The variation of verb choice to describe the variety of action being reported can also be seen, with verbs such as *jumping*, *carried*, *gunking*, *draw*, *circling*, *redeem*, and *cleaned*. Simply put, there is a lot of action to report, and the reporting of that action is likely to rely more on the use of verbs. Interestingly, while deviating from Biber et al.'s (1999) findings, these results align closely with the overall rate of occurrence of verbs in COCA, where verbs account for about 15.7% of words in the corpus (Davies, personal correspondence, April 29, 2017).

5.2 Other Parts of Speech

Counts of other parts of speech were also gathered, including prepositions/subordinating conjunctions, adjectives, modals, determiners, adverbs, pronouns, interjections, wh- words, coordinating conjunctions, existential *there*, foreign words, and cardinal numbers.

Part of speech	WiiM raw	WiiM frequency	COCA raw	COCA frequency

Preposition/sub. conj.	90582	11.1/100	68952303	12.9/100
Modal auxiliary	11237	1.4/100	5854104	1.1/100
Determiner	86180	10.6/100	15570233	2.9/100
Interjection	603	.07/100	951436	.18/100
Coordinating conjunction	24298	3/100	18451948	3.5/100
Existential <i>there</i>	1242	.15/100	1025805	.19/100
Foreign word	417	.05/100	200290	.04/100
Cardinal number	44581	5.5/100	3548536	.67/100

Table 4.3 Other parts of speech

Adjective/Ordinal Numeral	50753	6.2/100	38251943	7.2/100
Adjective, comparative	2590	.32/100	939332	.18/100
Adjective, superlative	1855	.23/100	559905	.11/100
Total	55198	6.8/100	39751180	7.5/100

Table 4.4 Adjectives

Adverb	40867	5/100	28603572	5.4/100
Adverb, comparative	1243	.15/100	833123	.16/100
Adverb, superlative	347	.04/100	93938	.02/100
Total	42457	5.2/100	29530633	5.5/100

Table 4.5 Adverbs

Personal pronoun	33601	4.1/100	29842393	5.6/100
Possessive pronoun	9245	1.1/100	8521306	1.6/100
Total	42846	5.3/100	38363699	7.2/100

Table 4.6 Pronouns

Wh- determiner	3159	.39/100	N/a*	N/a*
Wh- pronoun	3570	.44/100	N/a*	N/a*
Possessive wh- pronoun	35	.004/100	N/a*	N/a*
Wh- adverb	3999	.49/100	N/a*	N/a*

Total	10763	1.3/100	N/a*	N/a*
-------	-------	---------	------	------

Table 4.7 Wh- words

*Data for wh-words was not provided for COCA.

A. Foreign words

As in the 2014 study, upon close examination, the foreign words category was disregarded. Many of the instances marked as foreign words were occurrences of the unpunctuated shortened form of versus, *vs*, and others were words such as *etc*, *cam* in the term *ref cam*, a shortening for *camera*, the term *em* as a clipped version of *them*, and occasionally foreign players' names, though most foreign names were correctly categorized as nouns. The term *meme* was also categorized as a foreign word. The clear conclusion is that, once again, the majority of the words in this category were erroneously categorized due to their nonstandard structure, and this number is not truly representative of the use of foreign words in the corpus.

B. Prepositions/subordinating conjunctions and determiners

Of the remaining tagged categories, prepositions/subordinating conjunctions and determiners were again the third- and fourth-most common word categories following nouns and verbs, consistent with the 2014 findings. Each of these categories once again accounted for around 1/10th of the total words in the corpus. The rate of occurrence for prepositions is comparable to Biber et al.'s (1999) reported finding for news and slightly below the rate found in academic prose, but notably higher than rates found in fiction and conversation, explained perhaps by WiiM's often informative or reporting purpose and the heavy use of nouns, as prepositions often take noun phrases as complements. The preposition rate was only

slightly lower than COCA's 12.9%, while the rate of determiners was significantly higher than the 2.9% rate found in that corpus (Davies, personal correspondence, April 29, 2017). The high rate of determiners is likely partly attributable to the high rate of nouns, as there is generally a positive relationship between these rates. Biber et al. (1999) found that the use of determiners tended to occur at a higher rate in academic prose and news reports, both information-dense types of text. A similar argument can be made for WiiM, with many posts being informative in nature.

C. Adjectives and adverbs

Adjectives account for 6.8% of total words, only slightly higher than the 6% rate found in the 2014 study, and adverbs account for 5.2%, about the same rate as found previously. In both cases, comparatives and superlatives accounted for only a small percentage of the total category. While Biber et al. (1999) found a close correspondence between the ratio of adjectives to adverbs and that of nouns to verbs, that correspondence is not reflected here in the WiiM corpus. Biber et al. (1999) also found adjectives to be more common than adverbs in the written news and academic prose registers, while the opposite was true in conversation and fiction, and in this case WiiM, as a written register and as a less personal and more informative register, compares accordingly. As Biber et al. (1999) noted, "adjectives are frequently used to modify nouns, thus adding to the informational density of expository registers such as news and academic prose" (pp. 504). It is thus unsurprising that WiiM shows a higher rate of adjectives than adverbs, but somewhat surprising that the ratio of the two does not correspond more closely to that between nouns and verbs. Biber et al. (1999) also suggested that the reason adverbs are more common in conversation and fiction is that "adverbs occur most commonly as

clause elements (adverbials) and thus co-occur with lexical verbs in adding information to the relatively short (and therefore frequent) clauses of conversation and fiction” (pp. 504). This can perhaps help explain the close ratio between adjectives and adverbs in the WiiM corpus. While WiiM is largely an informative register, there is a significant assumption of shared knowledge and the writing style is less formal than that likely seen in academic prose or even news pieces. This may lead to WiiM landing between academic/news and conversation/fiction in terms of length and complexity of clauses. Biber et al. (1999) also found comparative adjectives to be about twice as frequent as superlative, another ratio not reflected in the WiiM corpus, where comparatives appeared to be only about 50% more common. Regarding comparative and superlative adverbs, they found that these forms are less common with adverbs than with adjectives, and that comparative adverbs are more common than superlative adverbs. Both findings were also seen in the results for the WiiM corpus.

D. Personal and possessive pronouns

Personal and possessive pronouns combined occurred in the WiiM corpus at a rate of 5.2%, about the same rate as adverbs. This was well below Biber et al.’s (1999) findings in conversation and fiction, but slightly higher than the rates found in news and academic prose. This is another finding which likely reflects the blog’s hybrid purpose, with many posts designed to inform but some also to express opinion or personal stance. This is also reflected in the breakdown of nouns by person. Third person pronouns were by far the most common in the corpus. Including variations of *it*, third person pronouns occurred about two and a half times as often as first-person pronouns and about five times as often as second person

pronouns. The high rate of third person pronouns is likely related to the largely informative purpose of the blog. First person pronouns occurring at the second-highest rate is likely related to the secondary opinion/personal stance expression purpose of the blog. Second person pronouns are suggestive of interactivity, which is difficult to achieve in the blog posts themselves and is a more common feature of text found in the comments section, which was not included in this corpus. In the case of both first and third person pronouns, singular pronouns were, overall, more frequent than plural pronouns. However, while first person pronouns were split at about 60/40, in the case of the third person, singular pronouns were about three times as common. Even after removing *it*, the rate was about two to one. This may be topic-driven, with individual players and staff and their actions frequently referenced.

Biber et al. (1999) found masculine pronouns to be more common than their feminine counterparts across all registers, and this was the case with the WiiM corpus as well. In fact, feminine pronouns occurred at a miniscule rate relative to masculine, with masculine pronouns occurring about 170 times as often. While Biber et al. (1999) did not provide exact numbers for comparison, it is likely that the difference across their registers was not as stark as was found in WiiM. In the case of WiiM, the difference is likely largely attributable to the blog's primary topic. Simply put, all NHL players are male, all NHL coaches and GMs are male, and the vast majority of other NHL employees, hockey writers, and other relevant figures are male. Even most of the bloggers are male. Prospects and potential draft picks are male, players in other leagues are male, and while a female international hockey league does exist, it is rarely discussed in comparison with male international

hockey. Considering all of this, while the contrast is striking, it also has a logical explanation.

E. Modals

Modals occurred at a rate of about 1.4%, again landing between conversation/fiction and news/academic prose, per Biber et al. (1999). This once again lines up with WiiM's hybrid purpose, as according to Biber et al. (1999), modals occur at the highest frequency in conversation because they "mostly convey stance-type meanings" (pp. 487). The WiiM corpus showed, relative to Biber et al.'s (1999) findings, slightly higher rates of *will*, *can*, and *should*, essentially the same rates of *could* and *might*, and slightly lower rates of *would* and *may*. *Must* occurred at a lower rate, while *shall* occurred at a significantly lower rate, presenting only 26 occurrences in the entire corpus. The much less frequent occurrence of *shall* is the only truly notable deviance from Biber et al.'s (1999) average findings. This may well be due to the informative but relatively informal nature of the blog, and this is supported by Biber et al.'s (1999) finding that the lowest occurrence for this modal was in news text. They also found that *shall* not only occurred with less frequency than the other main modals examined in the WiiM corpus, but also that the rate of usage for this modal in American English, as opposed to British English, is miniscule, suggesting the term is nearing archaic in this dialect. WiiM is a blog based in the United States with American bloggers discussing an American sports team, so this dialectally-driven lack of usage across American English registers is likely also a contributing factor.

F. Wh-words

Wh-words occurred in the WiiM corpus at a similarly low rate to modals, at an overall rate of about 1.3%. Little data was provided by Biber et al. (1999) and COCA rates for these words were unavailable (Davies, personal correspondence, April 29, 2017). However, interrogative sentences are significantly more common in conversation than in the other registers studied by Biber et al., with fiction coming in a very distant second, and an examination of WiiM posts suggests that questions are relatively uncommon in that text as well. This is supported by an examination of the use of punctuation, as only 2,375 question marks were found, despite a sentence count of almost 40,000.

G. Coordinating conjunctions

Coordinators occurred at a rate of about 3% in the WiiM corpus. This rate falls in between the rates Biber et al. (1999) found in fiction and academic prose, where the rates were higher, and conversation and news, where they were slightly lower. *And* is significantly more common in the WiiM corpus than *or*, *but*, and *nor*, just as Biber et al. (1999) found across the registers they examined. They found that *and* occurred more frequently in fiction and academic prose than in conversation and news, and WiiM also falls in between these two sets, though the rate is closer to the lower end with conversation and news. They suggested the high rate found in academic prose to be due to the use of *and* as both a phrase-level and clause-level coordinator and that its low rate in conversation reflects the heavy use of verbs requiring clause-level connection, which may be better achieved with *but* and subordinators. This may explain the WiiM corpus rate falling between the two but closer to the lower end, as WiiM likely uses less complexity and thus less phrasal

coordination than academic prose. Furthermore, WiiM showed a higher rate of verb use than average but also a high rate of nouns to verbs, so if the use of *and* correlates negatively with the heavier use of verbs for the reasons suggested by Biber et al. (1999), a resulting middling rate of *and* in the WiiM corpus would follow that reasoning. The rate of occurrence for *or* was above that found in news, below that found in academic prose, and close to the same as found in fiction and conversation. Biber et al. (1999) suggested the higher occurrence of *or* in academic prose was related to alternative explanations as well as the need to explain terms, and these purposes are not characteristic of WiiM thanks to the reporting nature of the informative purpose and the general expectation of shared knowledge. Biber et al. (1999) found the rate of use of *but* to be very low in academic prose, slightly higher in news, and the highest, at about the same rate, in fiction and conversation. The rate found in the WiiM corpus of about .57% is close to that found in fiction and conversation and is thus on the higher end of the spectrum of Biber et al. (1999) findings. They suggested that the higher rates found in conversation and fiction were likely related to a higher frequency of negation and contrast and thus reflect interactivity, a characteristic that does not strongly fit WiiM's blog posts, as there is only one participant. Thus, this high occurrence is not explained by the potential reasons offered by Biber et al. (1999). However, they also suggested the low level in academic prose is due to a preference for other, more formal forms of expressing contrast, such as *although* and *nevertheless*, and this preference is not likely to exist in a relatively informal setting such as sports blog posts. *Nor* was found to be so rare in the Biber et al. (1999) corpora that it was left out of the data graphic. In the WiiM corpus this word was used 36 times for a rate of .004%, also quite rare, though it is worth noting that it did occur and more than once. Biber et al. (1999) noted that

“negation is less frequent overall than positive forms” and that there exists a “preference for negation by *not* in conversation” as an explanation for the low distribution of *nor* in their corpora and its higher occurrence in fiction than in other registers, and the WiiM text likely mirrors both of these conditions (pp. 82).

H. Cardinal numbers

Cardinal numbers occurred in the WiiM corpus at the unusually high rate of 5.5%. Biber et al.’s (1999) highest rate found was around 2.2%, in news, with academic prose only slightly lower. This means that the WiiM corpus showed almost two and a half times as many cardinal numbers as the highest-frequency register covered by Biber et al. (1999), a somewhat striking finding. However, this is again likely explained by the combination of the heavy focus on topic and the informative and reporting purposes of WiiM. WiiM posts discuss a sport and many of the posts report game events, which often involve cardinal numbers, including numbers of goals, assists, penalties, players, et cetera, as well as numbers relating to time. Furthermore, even many opinion and personal stance pieces feature a notable use of cardinal numbers, as a variety of statistical information as well as time-related numbers such as counts of days, weeks, months, or years may be part of the discussion. The unusually heavy use of cardinal numbers can be seen in this excerpt from a post-game recap in which cardinal numbers are used 10 times in just 118 words:

6 minutes in Alexey Marchenko takes a hooking penalty and 12 seconds into the penalty Vincent Trocheck centers the puck to Colton Sceviour and he taps it in past Mrazek. Panthers up 1-0.

Sceviour finds himself in front of the net and buries the puck past Mrazek just 23 seconds after the power play goal and the Panthers are up by 2.

Red Wings get a power play on an Alex Petrovic interference. Nothing happens with for the Wings, but Andreas Athanasiou received some time on the power play with Sheahan and Larkin.

Florida breaks in 2v1 and a brilliant saucer pass from Denis Malgin finds Jonathan Marchessault who buries it over the shoulder of Mrazek. 3-0 Panthers.

I. Existential *there*

Existential *there* occurred at a rate of .15/100, very close to the .16/100 rate found in the 2014 study. This is only slightly below the rate found in COCA of .19/100 (Davies, personal correspondence, April 29, 2017). Both results are below Biber et al.'s (1999) results ranging from .2 in news to .3 in conversation and fiction, with academic prose falling in between. With the informative purpose of WiiM, it is not surprising that the corpus would fall closer to news text than to conversation or fiction in the usage of existential *there*, yet it is surprising that the rate of occurrence would be notably below all registers. Biber et al. (1999) suggested that existential *there* is most commonly used not just to introduce new elements in discourse but "to focus on the existence or occurrence of something" and is thus "most typically used with indefinite notional subjects" (pp. 951). The definite article *the*, accounting for about 6% of all words, was far more common than its indefinite counterparts *a* and *an*, accounting for about 2.7% together, and the same reasoning may apply to this occurrence as well. As previously discussed, there is a significant assumption of shared information in the WiiM posts, and that likely lends itself to information being introduced as new only infrequently.

J. Interjections

Interjections were uncommon in the WiiM corpus, at a rate of only about

.07/100. The rate found in COCA was notably higher, at around .18/100. According to Biber et al. (1999), interjections “generally operate at an emotive level of communication where nothing in the form of a proposition need be implied” (pp. 1104) and WiiM has not shown to be a highly emotive register. Furthermore, Biber et al. (1999) primarily discussed interjections in the context of spoken language and text derived from spoken language. As WiiM is a written register and has not shown to be the type of written register that showcases characteristics and features similar to spoken registers as one may expect from more interactive written registers such as chats and SMS text messages, a low rate of occurrence of a word category commonly considered to be a spoken register characteristic is to be expected.

6. Lexical Information

According to Antconc’s word counts, the WiiM corpus contained 20,706 types and 813,435 tokens. Of the top ten most frequent words, almost all were common function words. The only exception was the lexical word *wings*. That this lexical item would rank as high as seventh overall and fall in the top ten with function words speaks to the topic-driven nature of WiiM.

Top 10 words	Total number of occurrences
the	49284
to	21258
a	19403
and	16680

in	14360
of	14219
wings	8111
for	7951
on	7946
that	7643

Table 4.8 Top 10 words

When function words are removed, the top words appear to be very topic-specific.

Top 10 non-function words	Total number of occurrences
wings	8111
red	5353
game	4546
team	3453
season	2723
play	2583
detroit	2186
nhl	2142
time	1937

goal	1837
------	------

Table 4.9 Top 10 non-function words

All but one of these words have clear semantic connections to the topic of hockey, and several specifically to the Detroit Red Wings team. The only exception is *time*, though this likely also has a topic-specific explanation, as for example the amount of time left in a period, the amount of ice time a player played, and the amount of penalty time given are all common points of discussion, and the hockey-specific term *to one-time (the puck)* uses the word as well. It is notable that the term *red* occurs less frequently than *wings*. An inspection of the corpus shows that the team is often referred to simply as *the Wings*, forgoing the longer name *the Red Wings*.

7. Biber's MAT analysis

Finally, the WiiM corpus was analyzed using Biber's MAT analysis platform, which analyzes and aligns text relative to other registers along six different register-driven dimensions. Analyzing the corpus using MAT offers additional information on WiiM posts as a text type as well as situating the blog among other registers and descriptive clusters. MAT relies on the Stanford Tagger for its tagging as well, helping to maintain comparability of results.

7.1 Dimension 1

The WiiM corpus was analyzed across all six dimensions. Dimension 1 analyzes text on a spectrum of involved to informational. The analysis showed the WiiM text falling on the informational side of this spectrum, with its closest associated register being academic prose. The text was also close to press reportage text on this dimension. It was less informational in nature than official documents,

but more informational and less involved than general fiction, personal letters, prepared speeches, broadcasts, and conversations. Conversations was the register furthest from WiiM in this dimension. This result is unsurprising based on the previous discussion of how the analysis of WiiM's linguistic features compared to Biber et al.'s (1999) analysis of conversation, fiction, academic prose, and news and the previously-established purpose of WiiM being largely informative in nature.

7.2 Dimension 2

Dimension 2 analyzes text on a spectrum of narrative to non-narrative. The WiiM text fell on the non-narrative side of this spectrum, a result that speaks against the early definition of blogs as a sort of online personal diary. On this dimension, the WiiM corpus was once again most closely associated with academic prose. Official documents and broadcasts showed a somewhat close relationship as well. Personal letters and especially general fiction were significantly more narrative than the WiiM text, while conversation, prepared speech, and press reportage also fell higher on the narrative end of the spectrum. This is a somewhat surprising result given the WiiM corpus' high occurrence of past tense verbs and third person pronouns, both features Biber and Conrad (2001) associated with the more narrative side of the spectrum.

7.3 Dimension 3

Dimension 3 analyzes text on a spectrum of explicit to situation-dependent. The analysis on this dimension showed the WiiM corpus being very close to the middle of the spectrum, but slightly to the context-independent side. Prepared speeches was the register given as the closest, but press reportage was very close to WiiM on this spectrum as well. Official documents and academic prose were both

notably less context-dependent, while general fiction, personal letters, conversations, and especially broadcasts were all much more context-dependent. While an analysis of a corpus built of comments from the WiiM game threads would likely show a much more heavily context-dependent result, this more neutral result for the blog posts themselves is unsurprising. While a significant amount of shared knowledge is assumed across the blog, many WiiM posts report game events or other events involving the team or the league, and this purpose does not necessarily demand an awareness of the immediate context (e.g., one does not have to be watching the game while reading game summaries in order for the game summary posts to successfully achieve their purpose of informing the reader of game events). A close-to-neutral result reflects this combination of shared knowledge expectancy and reporting and informative purpose. Biber and Conrad (2001) suggested that a higher rate of adverbs is a negative feature on this dimension, meaning that the more positive the score- the more explicit the text- the higher the rate of adverbs, and indeed in analyzing linguistic features for WiiM, the occurrence of adverbs was somewhat high relative to the occurrence of verbs, which would be expected to correlate more closely.

7.4 Dimension 4

Dimension 4 analyzes text on a spectrum of overt expression of persuasion. WiiM fell slightly to the negative side of this spectrum, below the means of most of the comparison register. Only broadcasts had a mean lower than the WiiM result. Press reportage was the given closest register. While based on the aforementioned early definition of blogs as places of expression of personal thoughts and opinions this would be a surprising result, based on the discussions of WiiM's characteristics

and linguistic features and functions, this is not unexpected. Some blog posts are of a somewhat editorial nature, but the primary purpose of WiiM as a blog appears to be informative and reporting, and considering that, a result on this dimension close to that of press reportage is unsurprising. Official documents, academic prose, and conversation were also somewhat close to the WiiM result, while prepared speeches and general fiction were higher and personal letters much higher on the overt expression of persuasion spectrum. WiiM's rate of occurrence of modals, landing between conversation/fiction and news/academic prose, was likely a factor in this result.

7.5 Dimension 5

Dimension 5 analyzes text on a spectrum of abstract to non-abstract information. The results of the WiiM text fell slightly below neutral on the non-abstract side, with the closest register being broadcasts. Press reportage also had a mean somewhat close to the WiiM result, though slightly above neutral on the abstract side. Official documents and academic prose were much higher on the abstract side, while general fiction, personal letters, prepared speeches, and conversation were much lower on the non-abstract side. According to Biber and Conrad (2001), passive construction was particularly commonly associated with abstract information on this dimension, and passive language is not especially common in the WiiM corpus. They also suggested that conjuncts such as *thus* and *however* were associated with abstract information, and neither of these words was common in the WiiM corpus, with *however* occurring just 329 times and *thus* just 33 times. In comparison, the coordinator *but*, similar in purpose to *however*, occurred over 4600 times and is thus the clearly-preferred construction. Per Biber and Conrad

(2001), this dimension is also sometimes referred to as impersonal vs non-impersonal style, and while WiiM may have a primarily informative purpose, personal stance, opinion, and informality are permissible in the general writing style of the blog, and special efforts to present information from an impersonal standpoint are not made such as they would be in academic prose or even press reportage.

7.6 Dimension 6

Dimension 6 analyzes text on a spectrum of on-line informational elaboration. The result for the WiiM corpus on this dimension was negative of neutral, with official documents being the closest register. Press reportage also showed a very similar result, and broadcasts and personal letters were somewhat close as well. General fiction was also on the negative end of the spectrum as the lowest result, with conversations, prepared speeches, and academic prose all positive. Biber and Conrad (2001) found that this dimension "seems to be associated with spoken registers that are informational in focus and that convey speaker attitudes and beliefs" (pp. 41) with the only register showing a drastically positive result in this dimension being prepared speeches. As WiiM is not a spoken register, a negative result in this dimension is unsurprising.

Chapter 5

The Authorship Study

1. Background

The history of authorship study as both a subfield of linguistics and a field in its own right extends back quite some time, with a century-and-a-half-long timeline of studies which contribute to the field as it is known today. Following is a brief overview of the major contributors and their techniques, approaches, and specialties, which together form the foundation of the field of authorship examination in forensic linguistics.

1.1 Early authorship work

Authorship studies fall primarily within the realm of the field of forensic linguistics. Forensic linguistics is a relatively young subfield of linguistic study covering disciplines at the intersection of the scientific study of language and some aspect of the law. Linguists can function as consulting experts regarding a variety of topics at this intersection, from proper linguistic handling of child witnesses and victims to determination of linguistic origin of asylum seekers in cases of refugee nationality claims. One prominent area in which linguists have become very involved is text analysis. Most frequently, such experts are sought to assist in the determination or verification of authors of texts.

Though the field of linguistic authorship examination is relatively young compared to other linguistic subfields, it finds its roots in studies from as early as the mid-19th century. The earliest documented foray into text analysis for the purposes of authorship authentication is the work of Augustus de Morgan, in 1851. De Morgan,

a mathematician, proposed in a letter that comparing average word lengths could potentially provide insight into the authorship of documents, specifically the Epistles of Paul (De Morgan, 1882). De Morgan never tested his theory, but his letter was published posthumously in 1882, and, notably, was read by geophysicist T.C. Mendenhall (Grieve, 2005).

Mendenhall, inspired by de Morgan's proposition, embarked on an extensive series of text examinations theorizing and testing out methods for authorship authentication. His experiments were built primarily on the foundation of average word length suggested in de Morgan's letter. Mendenhall focused on the distribution of an author's word-length frequency, which he termed the author's word-spectrum. This work led to his 1887 publication, *The Characteristic Curves of Composition*, in which he graphed the distribution of those frequencies (Mendenhall, 1887).

In reaction to Mendenhall's work, H.T. Eddy was compelled to expand on Mendenhall's ideas. In an 1887 letter, Eddy proposed that average sentence length and sentence length distribution may potentially offer more robust evidence of authorship (Eddy, 1887). In 1888, William Benjamin Smith, as Conrad Mascol, followed Eddy's work with his own experiments on average sentence length (Grieve, 2005). However, while Eddy examined sentence length in words, Smith looked at the number of sentences per page (Mascol, 1888).

In the 1930s, statistician G. Udny Yule applied his expertise by implementing statistical methods of text analysis, utilizing a variety of techniques. He, too, looked at sentence length distribution, as well as vocabulary richness measures (Yule 1939). Yule developed a statistical algorithm, termed Yule's Characteristic, for measuring and comparing vocabulary richness. He also looked at the frequencies of graphemes

as well as their distributions, theorizing, for example, that authors may have preferences for words beginning with a specific letter.

In the 60s, three researchers began extensive examinations of authorship attribution techniques, effectively kick starting the modern wave of authorship work. Mosteller and Wallace dismissed the idea of sentence length as a reliable authorship technique, suggesting that the method did not adequately distinguish among authors. Instead, they proposed that the use of function words may provide a more accurate determination of authorship. They theorized that function words were less likely to be influenced by the context of the writing, such as subject matter and audience, than measures such as sentence length. They also experimented with the frequency of nouns and adjectives and the frequency of one- and two-letter words. They tested their techniques out on the infamously-disputed Federalist Papers, and published *Inference and Disputed Authorship*, which remains a highly-influential and heavily-cited publication in the field of authorship studies (Mosteller & Wallace, 1964).

Around the same time that Mosteller and Wallace were performing their research, another prolific authorship scholar was also experimenting with a variety of techniques. Andrew Morton continued to examine the stability of sentence length distribution. Similarly to Mosteller and Wallace's work, he, too, experimented with function word distribution. Morton became interested in word position stylometry, and specifically examined the position of function words. He also examined collocations as a possible tool for determining authorship, including co-occurrence of certain function words with other words. In 1978, he published the book *Literary Detection* which outlined his experiments, methods, and results (Morton, 1978).

While Mosteller and Wallace and Morton contributed significantly to the foundations of modern authorship research and are frequently referenced, a handful of other scholars made important contributions over the course of the next decade as well. In 1966, Bernard O'Donnell performed some of the earliest work with syntactic methods. He examined texts for relative parts of speech frequency, as well as for such syntactic measures as the frequency of clauses, of dependent clauses, and of past participle sentences (O'Donnell, 1966). A couple of years later, in 1968, Svartik performed some of the earliest work on non-literary text- and, crucially- specifically for forensic purposes- when he examined alleged confessions of a murder suspect (Svartik, 1968). Svartik also used syntactic techniques, including the frequency of clauses. In 1975, Damereau suggested that Mosteller and Wallace's function words methods were questionable by proposing that function word distribution was not necessarily as random as previously supposed- that is to say, that function words are not used in the random manner suggested by the proposition that they are reliable methods of authorship determination (Damereau, 1975).

1.2 Modern Research

Over the last three decades, the field has seen a surge of linguistic research on authorship authentication. While some of this research has continued to be applied for literary uses, a much higher portion of the research than in the early stages of the field's development is carried out for forensic applications. Both areas of application provide a critical foundation for the research conducted for this dissertation.

The first scholar who warrants discussion here is Donald Foster, a literature professor at Vassar College. Foster utilized stylistic techniques to examine a wide

variety of text types, including both long literary works and relatively short personal letters. He conducted this work both for literary purposes and for forensic ones (Foster, 2002). However, both Foster's methods and his results have been widely criticized. It is difficult to even test his methods for replicability, both because he has published very little on them besides his 2002 book, which contains little methodological detail, and because many of those methods do not appear to adhere to the scientific method, making them difficult to reproduce even with detailed information. The linguistic authorship community appears to overwhelmingly view Foster as little more than a cautionary tale of how not to conduct one's research or comport oneself in the forensic linguistic setting.

The linguistics scholars who work on authorship authentication techniques can be divided into two broad categories: those who focus primarily on stylistic techniques, and those who focus on stylometric techniques (Grant, 2013). McMenamin is perhaps the most renowned of the scholars on the stylistic side of the spectrum. McMenamin has spent several decades using linguistic-based techniques to perform stylistic examinations of a variety of texts, including the Jon Benet Ramsey ransom letter (2002). McMenamin uses a variety of both qualitative- such as capitalization and punctuation habits- and quantitative- such as spelling and grammar errors- approaches to make his authorship determinations.

Carole Chaski approaches the problem of authorship authentication from the stylometric side. Chaski is a syntactician and focuses primarily on underlying syntactic structure, which she suggests is subconscious and thus difficult for the author to manipulate while also offering enough differentiation to allow for manifestation of author idiolect (Chaski, 2013). Chaski has developed several software platforms based on her theories involving syntactic structure which, ideally,

can be utilized by individuals who are not linguistic experts, such as investigators, to determine or confirm authorship (Chaski, 2007). Chaski strives toward establishing methods with well-documented minimal error rates and tools that will allow those needing these types of services who are not linguistic experts to simply input text as data and receive output in the form of probability of authorship. Chaski frequently discusses the problem of ground truth data in authorship work. She suggests that, while researchers are testing out methods to establish validity and error rates, the identities authors of the data used must be known or at least accessible to the researcher. Chaski further warns that, particularly in the case of using Internet-based data, it is absolutely paramount that the researcher is not just able to identify an author for a text, but that they can also confirm that authorship via existing, certain knowledge of the identity of the author (2013). This can be an issue particularly on public platforms with multiple contributors. The researcher must be able to confirm that only one individual posts under a specific username, for example, before collecting texts from that username for comparison. In opposition to McMenamin and other stylistics scholars, Chaski tends to eschew qualitative methods, preferring the objectivity of quantitative measures (Chaski, 2001).

The problem of authorship authentication has captured the attention of computational linguists as well as computer scientists who are not trained in linguistics at all. These two fields have also produced research on methods for authorship authentication. Their methods are almost exclusively automated, and many involve machine learning and programming algorithms. Patrick Juola (2012), Kim Luyckx together with Walter Daelemans (2011), David Holmes (1992), and Efstathios Stamatatos (2013), as well as the research team of Moshe Koppel, Jonathan Schler, and Shlomo Argamon (2009) have all produced extensive research

on these types and techniques and achieved some success. While their methodologies involve complex coding skills that are presently beyond the scope of this dissertation, these scholars have contributed significantly to the body of work on authorship authentication and this area of authorship work must be mentioned in any overview of the field.

Authorship authentication research has a plethora of applications both in literary analysis and in forensic settings. An application in which it has proven to be an extremely valuable service is the world of plagiarism detection. The body of authorship research conducted toward solving issues of plagiarism detection has become quite extensive. A variety of software platforms are now available on both the consumer and industry markets that will help detect potential plagiarism, alerting e.g., teachers and professors to potentially problematic cases for further review. David Woolls is a premiere researcher on authorship work for the purposes of plagiarism detection. He argues that computational means of detection provide a massive leap forward in the endeavor due to the ability of software to handle massive amounts of both suspect text and comparison text in a very short period of time (2012). This is especially valuable in that it allows instructors and institutions not only to handle checks of a large number of students' work, but also to compare that work against vast repositories of available comparative text. This includes conducting checks which scrape the internet for text with a high degree of similarity. Woolls has developed two programs, Abridge and Vocalyse Toolkit, which are designed to automate plagiarism detection for use by those who are not necessarily linguistic experts.

2. Experiment design

2.1 Methodology

A. Building author corpora

To begin the authorship authentication examination, I first set out to choose blog authors from WiiM for whom enough data was present to build corpora. I chose three main contributors to the blog: JJ From Kansas, KyleWiiM, and Jeff Hancock. JJ From Kansas is the longest-tenured blogger on the site, with several thousand archived posts. KyleWiiM and Jeff Hancock are both long-standing, frequent contributors with consistently high levels of administrative access.

After choosing which bloggers to work with, the next step was to explore their blog post archives to extract posts. I followed the common choice of taking 10 texts from each author, as done in studies by such authorship researchers as Chaski (1999) and Grieve (2005). I set a minimum word count of 300 for viable posts in order to ensure my ability to compile enough data for each author via the 10-post count. Because of this minimum, posts that contained little author-created text, such as Quick Hits posts, which largely consist of links and may include little to no author commentary, were not used for these author corpora. In addition to leaving out Quick Hits posts and other posts with fewer than 300 words of original author text, all texts chosen were examined closely for non-author contributions. Quotations of others' words, included Twitter posts, chunks of text directly copied from other sources, author bylines and other credit lines, and data in the form of infographics taken from other sources were deleted, as none of these types of text were original language from the authors. Any text that was merely part of the blog's formatting was also removed, as were all images, videos, gifs, and other media. Only text which

appeared to originate from the author himself was retained. This included parenthetical update annotations.

Once text that was not attributable to the blogger himself was removed, the text was otherwise left unaltered. Any potential typos or errors, including spacing errors, were left in place. As errors will be one parameter by which the documents are examined, these errors needed to be retained.

B. Choosing parameters

An initial list of possible parameters to examine was built based on previous research in the area of authorship authentication and verification. The comparative studies of Chaski (1999) and Grieve (2005) provided the primary inspiration for parameter options for this study. Parameters were chosen based on feasibility of accurate examination within the confines of this study. As the ability to code custom utilities and examine large amounts of complex data via computational methods was not available, parameters requiring access to such methods were not chosen for the current study. Once the list of parameters was compiled, the parameters were examined more closely alongside the chosen documents and available utilities to ascertain the feasibility of examining each parameter within the scope of this study.

C. Parameters to be examined

The first parameters chosen were comparisons of average word length and average sentence length. These parameters were further broken down into average sentence length in words and average sentence length in characters and also included comparisons of the lengths of the shortest and longest sentences in words. Word length profiles were also compiled for comparison. Type-token ratio was also

chosen as a parameter for comparison, as was lexical density and syllable count. Readability metrics were compared, including those based on the following formulas: the Flesch Reading Ease, the Flesch-Kincaid Grade Level, the Gunning FOG Scale, the SMOG Index, the Coleman-Liau Index, and the Automated Readability Index (ARI). The grapheme profiles of the test documents and corpora were also compared as a parameter. Parts of speech counts were compiled as a parameter for comparison, and usage of individual modals was also compared. Verb forms and their indicated usage were examined for comparison, as were function word profiles. Spelling, punctuation, and grammar errors, including apparent typos, were examined as parameters. Punctuation mark profiles were compiled and compared, and finally grapheme N-grams were compiled and compared as a parameter as well. Finally, in addition to simple punctuation mark profiles, syntactically-classified punctuation was examined.

D. Obtaining baseline measurements for author corpora

Each of the three author corpora was examined via the chosen parameters to set the baselines in each parameter for each author. The process began with running each of the author corpora through the wordcounter.net utility to obtain counts of average word length, average sentence length in words, and average sentence length in characters. Textalyser.net was then used with each set of data to obtain the length in words of the shortest and longest sentences for each author corpus. Running the corpus data through the Textalyser utility also produced results for word length profiles. The rates of each possible length of words in characters being recorded for comparison, and a table of the lengths ranked in order of frequency from most to least frequent was compiled. Textalyser also provided counts for words based on number of syllables. Wordcounter.net was also used to gain type and token counts for each corpus, which were then combined to determine type-token ratio.

Analyzemywriting.com's utility was used to determine lexical density, which is gained by dividing the number of lexical words in the corpus by the number of total words.

Readability metrics were determined using two separate utilities, webpagefx.com's Readability Test Tool and online-utility.org's Readability Calculator. Both utilities examine documents using the same indexes. However, as it was discovered that they examined and broke down aspects of the documents differently, leading to different results for those indexes, the data was run through both for comparison to examine whether the choice of utility could impact the results.

The corpora were run through dcode.fr's Frequency Analysis tool to determine character-level N-grams. Counts were obtained for unigrams, bigrams, and trigrams of all three author corpora. The author corpora were next run through the Stanford Tagger utility to obtain parts of speech counts. The tagged documents were run through Antconc to gain frequency counts of each tag, which were combined when necessary to determine final counts for the categories of noun, verb, adjective, adverb, determiner, existential *there*, modal auxiliary, coordinating conjunction, interjection, wh-word, and preposition or subordination conjunction.

Antconc was also used to obtain the top 20 function words for each corpus via the key word function. The corpora were then searched as Word documents for punctuation marks, including periods, commas, colons, semicolons, parentheses, quotation marks, ampersands, plus signs, hyphens, slashes, ellipses, question marks, apostrophes, and exclamation marks, and counts for each type of punctuation were recorded. Finally, the corpora were run through Lancaster University's online CLAWS tagging utility to tag specific types of verbs with the CLAWS7 tag set. These tags were then used to determine counts for infinitive,

passive, past participle, and -ing participle verbs. Each of these verb types was normalized per total number of verbs and per total number of words from each corpus for comparison purposes.

I examined each author corpus manually for errors for the purpose of comparing specific errors between the author corpora and the test documents in search of common errors indicating idiosyncratic patterns. Grammatical errors and spelling errors, as well as other errors such as punctuation and spacing, were noted and recorded. Errors which appeared to be systematic or habitual because they occurred multiple times and potentially in multiple contexts were noted as such.

The final parameter was syntactically-classified punctuation, which I also examined manually. Punctuation marks were classified for the following types of usage: abbreviation periods; list periods; decimal periods; end-of-sentence periods; end-of-sentence question marks; end-of-sentence exclamation points; quotation marks on sentences; quotation marks on words; quotation marks on phrases; apostrophes used for contractions; apostrophes used for plurality; apostrophes used for possession; list commas; commas separating main clauses; commas separating main and dependent clauses; commas separating phrases; semicolons in lists; semicolons separating main clauses; hyphens within words; hyphens between main clauses; hyphens between main and subordinate clauses.

E. Compiling and obtaining measurements for test documents

Once author corpora had been compiled and examined for parameter data, test documents were obtained from WiiM's archives. The same constraints as were used to choose posts for the author corpora were used for choosing these test documents, namely that the chosen blog posts had to be at least 300 words of the

author's own writing. One blog post was chosen as a test document from each of the authors included in the corpora, resulting in three total test documents, one written by each of the three authors (Jeff, JJ, and Kyle). The documents were coded with numbers for file names to keep the identity of the author of each document hidden during the study. Posts were chosen that were not already part of each author's corpus, but as close to the time period of posts culled for the corpora as possible to help control for possible temporal changes in the author's language habits. Thus, posts were chosen which were posted either shortly after or shortly before the time period from which the corpus posts were taken. The criteria for these posts were that they were written by one of the three bloggers, that they had a minimum of 300 words of the author's own writing, that they were not included in the corpus already, and that they were as close as possible to the time period from which the corpus posts were taken. Having to adhere to these criteria strictly while choosing test documents meant that I as the researcher needed to make the selections, and this made it impossible to maintain absolute anonymity in the selection. However, this issue was offset by two factors. The first is that I chose the documents long before working on them, did not read them while selecting them, and immediately saved them with coded file names. The second factor was that, should any lingering knowledge of the identity of the authors remain, the quantitative nature of the methodology chosen for this particular authorship analysis meant that impact of bias would not be possible, as the numbers and their comparisons could not be inadvertently manipulated. A study of this kind relying on qualitative methods would require complete blindness to author identity on the part of the researcher, and this would have to be considered in structuring the study. As with the posts collected for the author corpora, each of these documents was edited to remove text that was not written by the author himself, including any embedded tweets as well as author and

credit bylines. Tables were kept if they were created by the author rather than embedded from another source.

Once the documents were obtained, they were each individually subject to the same processes as the author corpora discussed above. They were each run through all of the same utilities and the same counts were taken. Once all of the necessary data was obtained, cross-comparisons between each test document and each author corpus using each parameter were carried out. In order to carry out these cross-comparisons, when necessary (i.e., when the results were not already averaged as part of the parameter and thus could not be accurately compared across documents of different lengths), the counts taken from each parameter for each document or corpus were normalized to either a count per 100 or a count per 1000 words, obtaining, in essence, a percentage to allow for accurate comparison across texts with unequal word counts. Once normalized counts were obtained for parameters requiring them, I compared each test document's normalized-count results against the normalized counts obtained from all three author corpora for each parameter. Whichever author corpus had a normalized count for that parameter that was closest in number to the normalized count found for the test document was assigned as the author for that test document via that parameter. When more than one author corpus was equally close in number, that document was recorded as being assigned to both/all authors, as this is a case where the parameter was clearly unable to differentiate between two or even all three authors. Thus, for example, in the case of the parameter average word length, Document 01, with an average word length of 5.2, was closest to both Jeff and JJ's author corpora, which both showed average word lengths of 4.7, while Kyle's showed a lower result of 4.6. In this case, both JJ and Jeff were assigned as likely authors of Document 01 for the average

word length parameter. Document 02, with a result of 4.6, was assigned to Kyle, whose result was the same number and thus the closest. Document 03, with a result of 4.5, was also assigned to Kyle, whose result was the closest to that number with only a .1 difference, compared to the .2 difference between Document 03 and Jeff and JJ's respective 4.7 results. Tables were created to track the assigned author for each document via each parameter. After cross-comparison examinations were carried out and their results obtained and recorded, the true identity of the author of each test document was checked against those results for efficacy of the parameter in identifying the author correctly.

3. Results

Below, the results of the parameter examinations are recorded in tabular format. Where possible, the data are presented in the form of raw counts and/or normalized counts or percentages. In the case of spelling errors, actual errors are recorded for possible comparison.

3.1 Average sentence and word length

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
Avg. word length (characters)	4.7	4.7	4.6	5.2	4.6	4.5
Avg. sentence length (words)	20	36	20	24	27	12
Avg. sentence length (characters)	110	198	111	144	149	65

Shortest sentence (words)	1	1	1	1	5	1
Longest sentence (words)	62	139	59	48	41	32

Table 5.1 Average sentence and word length

3.2 Type-Token Ratio

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
Types	1793	2679	1938	243	236	318
Tokens	5279	9018	5835	420	401	607
Ratio	.34	.297	.33	.58	.59	.52
Lexical Density	56.6	53.6	52.64	54.91	51.76	51.81

Table 5.2 Type-token ratio

3.3 Readability- Readability Test Tool

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
Flesch Reading Ease	66.9	62	67.1	65.1	63.6	83.2
Flesch-Kincaid Grade Level	8.4	9.9	8.7	9.5	11.4	4.3
Gunning-Fog	9.9	11.7	11.1	10.4	13.6	6.7
SMOG	7.9	9	8.3	8	8.7	5.3
Coleman-Liau	9.9	9.3	9.6	8.5	8.7	8.4
ARI	8.1	9.4	8.6	8.7	12.3	3.3

Table 5.3 Readability- Readability Test Tool

3.4 Readability- Online Utility

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
--	------	----	------	-------	-------	-------

Flesch Reading Ease	63.21	58.97	64.69	59.83	57.46	79.95
Flesch-Kincaid Grade Level	8.97	10.49	9.24	10.33	12.35	4.79
Gunning-Fog	11.02	12.79	11.66	11.96	14.36	7.12
SMOG	10.84	11.90	11.04	11.11	11.94	8.13
Coleman-Liau	8.61	8.45	7.63	9.10	7.53	6.45
ARI	8.52	10.04	8.57	10.48	12.27	3.98

Table 5.4 Readability- Online Utility

3.5 Parts of Speech

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
Nouns	1,904 36%	3144 33.4%	1854 30.1%	165 39.3%	112 27.3%	170 28%
Verbs	604 11.4%	1173 12.5%	856 14.2%	36 8.6%	71 17.3%	115 19%
Adjectives	432 8.2%	612 6.5%	499 8.3%	32 7.6%	29 7.1%	41 6.8%
Adverbs	228 4.3%	479 5.1%	264 4.4%	9 2.1%	19 4.6%	66 10.9%
Determiners	585 11.1%	942 10%	640 10.7%	30 7.1%	38 9.3%	66 10.9%
Existential There	5 .1%	6 .06%	6 .1%	0 0%	2 .49%	5 .82%
Modal Auxiliaries	43 .81%	146 1.6%	6 .1%	1 .24%	4 .98%	8 1.3%
Coordinating Conjunctions	142 2.7%	282 3%	131 2.2%	17 4.1%	12 2.9%	19 3.1%

Interjections	4 .08%	0 0%	0 0%	0 0%	0 0%	2 .33%
Wh-words	45 .85%	109 1.2%	110 1.8%	0 0%	4 .98%	9 1.5%
Prepositions or subordinating conjunctions	541 10.2%	990 10.5%	691 11.5%	58 13.8%	60 14.6%	58 9.6%

Table 5.5 Parts of speech

3.6 Modals

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
would	15 .28%	22 .24%	26 .45%	0 0%	1 .25%	1 .28%
should	2 .04%	14 .16%	9 .15%	0 0%	0 0%	0 0%
shall	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%
will	23 .44%	41 .45%	22 .38	0 0%	2 .5%	3 .84%
could	8 .15%	12 .13%	16 .27%	0 0%	0 0%	2 .56%
can	6 .11%	45 .5	26 .45	1 .24%	0 0%	1 .28%
may	0 0%	3 .03%	2 .03%	0 0%	1 .25%	0 0%
must	0 0%	0 0%	2 .03%	0 0%	0 0%	0 0%
might	1	11	0	0	0	1

	.02%	.12%	0%	0%	0%	.28%
--	------	------	----	----	----	------

Table 5.6 Modals

3.7 Function word profile

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
1	the	the	the	the	the	the
2	a	to	a	in	to	he
3	to	a	to	and	a	to
4	and	and	in	with	in	s
5	of	of	of	s	of	t
6	in	in	s	his	and	and
7	s	that	he	to	as	that
8	at	s	is	he	s	a
9	for	for	with	on	he	in
10	on	on	and	among	with	not
11	that	is	be	for	for	on
12	he	be	I	from	his	I
13	it	it	that	a	over	be
14	is	but	for	as	up	of
15	was	with	on	at	about	that
16	from	this	at	has	another	but
17	be	have	it	most	are	for
18	their	at	but	of	at	is
19	I	I	his	or	has	this
20	this	he	an	overall	have	was

Table 5.7 Function word profiles

3.8 Word length profile- length frequency counts and percentages

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
1	279 5.3%	589 6.3%	287 4.8%	12 2.7%	14 4.1%	59 9.1%
2	782 14.7%	1589 16.9%	954 16%	74 16.4%	75 18%	109 16.8%
3	1116 21%	1895 20.2%	1212 20.3%	92 20.4%	94 22.6%	141 21.8%
4	879 16.6%	1658 17.7%	1084 18.1%	61 13.6%	83 20%	132 20.4%
5	643 12.1%	1061 11.1%	739 12.4%	59 13.1%	38 9.1%	71 11%
6	516 9.7%	745 7.9%	552 9.2%	66 14.7%	36 8.7%	45 6.9%
7	404 7.6%	672 7.2%	452 7.6%	41 9.1%	23 5.5%	40 6.2%
8	265 5%	443 4.7%	272 4.6%	23 5.1%	21 5%	32 4.9%
9	153 2.9%	255 2.7%	186 3.1%	5 1.1%	16 3.8%	5 .8%
10	91 1.7%	213 2.3%	132 2.2%	5 1.1%	6 1.4%	13 2%
11	66 1.2%	123 1.3%	51 .9%	8 1.8%	4 1%	-
12	48 .9%	63 .7%	31 .5%	1 .2%	1 .2%	1 .2%
13	20	21	11	1	2	-

	.4%	.2%	.2%	.2%	.5%	
14	12 .2%	14 .1%	5 .1%	1 .2%	-	-
15	6 .1%	8 .1%	2 .03%	-	-	-
16	7 .1%	4	2 .03%	-	-	-
17	5 .1%	8 .1%	2 .03%	-	-	-
18	6 .1%	7 .1%	1 .02%	-	-	-
19	5 .1%	1 .01%	-	1 .2%	-	-
20	3 .1%	2 .02%	-	-	-	-
21	2 .04%	2 .02%	-	-	-	-
22	-	1 .01%	-	-	-	-
25	-	1 .01%	-	-	-	-
29	2 .04%	-	-	-	-	-
35	1 .02%	-	-	-	-	-

Table 5.8 Word length profiles- frequency counts

3.9 Word length profile- frequency ranking

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
1	3	3	3	3	3	3
2	4	4	4	2	4	4
3	2	2	2	6	2	2
4	5	5	5	4	5	5
5	6	6	6	5	6	1
6	7	7	7	7	7	6
7	1	1	1	8	8	7
8	8	8	8	1	1	8
9	9	9	9	11	9	10
10	10	10	10	10	10	9
11	11	11	11	9	11	12
12	12	12	12	19	13	-
13	13	13	13	14	12	-
14	14	14	14	12	-	-
15	16	17	15	13	-	-

16	18	15	16	-	-	-
17	15	18	17	-	-	-
18	17	16	18	-	-	-
19	19	20	-	-	-	-
20	20	21	-	-	-	-
21	21	22	-	-	-	-
22	29	19	-	-	-	-
23	35	25	-	-	-	-

Table 5.9 Word length profiles, ranking

3.10 Punctuation mark profile- raw count and normalized per 100 words

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
Period	284 .054	410 .046	294 .05	21 .05	14 .04	59 .097
Comma	196 .037	329 .037	277 .048	26 .06	17 .04	14 .023
Colon	53 .01	101 .011	41 .007	0 0	2 .005	3 .005
Semicolon	7 .001	7 .0008	5 .0009	0 0	0 0	0 0
Parentheses	38 .007	130 .014	17 .003	21 .05	2 .005	1 .002
Quotations	9	26	14	0	0	0

	.002	.003	.002	0	0	0
Ampersand	5	0	0	0	0	0
	.001	0	0	0	0	0
Plus sign	20	2	5	0	0	0
	.004	.0002	.0009	0	0	0
Hyphen	144	257	161	0	15	6
	.027	.029	.028	0	.037	.01
Slash	7	25	3	0	0	2
	.001	.0028	.0005	0	0	.003
Ellipses	6	4	1	0	0	5
	.001	.0004	.0002	0	0	.008
Question mark	6	10	9	0	1	2
	.001	.001	.0015	0	.003	.003
Apostrophe	138	226	192	0	10	35
	.026	.025	.033	0	.03	.058
Exclamation point	5	5	1	8	0	1
	.001	.0006	.0002	.019	0	.002

Table 5.10 Punctuation mark profiles

3.11 Verb forms- raw counts and counts per 100 verbs/words

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
Infinitive	128	341	230	3	16	28
Passive	36	88	42	1	3	4
Past participle	84	166	100	7	9	9
-ing participle	69	161	144	6	6	11
Infinitive/verbs	21.2	29.1	27	8.3	22.5	24.3

Infinitive/ words	2.42	3.78	3.94	.71	3.9	4.6
Passives/ verbs	6	7.5	4.9	2.8	4.2	3.5
Passives/ words	.682	.976	.72	.24	.73	.66
Past part/ verbs	13.9	14.2	11.7	19.4	12.7	7.8
Past part/ words	1.59	1.84	1.71	1.7	2.2	1.5
-ing part/ verbs	11.4	13.7	16.8	16.7	8.5	9.6
-ing part/ words	1.31	1.79	2.47	1.4	1.5	1.8

Table 5.11 Verb forms

3.12 Syllable counts, percentage relative to total word count

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
1	2831 58.5%	5021 59.7%	3482 62.5%	226 59.3%	238 62%	407 65.8%
2	1369 28.3%	2159 25.7%	1379 24.8%	104 27.3%	105 27.3%	154 25.3%
3	446 9.2%	835 9.9%	501 9%	36 9.4%	32 8.3%	42 6.9%
4	150 3.1%	324 3.9%	173 3.1%	9 2.4%	9 2.3%	11 1.8%
5	34 .7%	54 .6%	32 .6%	5 1.3%	-	1 .2%
6	6 .1%	12 .1%	-	1 .3%	-	-
7	1	1	-	-	-	-

	.02%	.01%				
8	-	-	-	-	-	-
9	1 .02%	-	-	-	-	-
10	2 .04%	-	-	-	-	-

Table 5.12 Syllable counts

3.13 Grammatical errors, raw count and normalized to per 100 words

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
Sentence fragment	3 .06	2 .02	10 .17	-	-	3 .05
Run-on sentence	5 .1	15 .17	11 .19	1 .24	1 .25	1 .17
Subject-verb mismatch	1 .02	3 .03	2 .03	-	-	-
Tense shift	2 .04	5 .06	5 .09	1 .24	-	-
Wrong verb form	11 .21	7 .08	11 .19	-	1 .25	-
Missing auxiliary verb	1 .02	-	-	-	-	-
Total	23 .44	32 .36	39 .67	2 .48	2 .5	4 .66

Table 5.13 Grammatical errors

3.14 Spelling errors

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
--	------	----	------	-------	-------	-------

Errors	Awhile	Can not	In regards to	-	-	Non
	Quite possible	Pricetag	Further adieu			their
	Stake	In regards to	Complimentary			
	Core	Goalscoring	Defenseman			
		Expecations	Onto			
		Inwards				

Table 5.14 Spelling errors

3.15 Syntactically classified punctuation- raw counts

*Notes: EOS- end of sentence; ?- question mark; !- exclamation point; main/sub- between a main clause and a subordinate clause; main/dep- between a main clause and a dependent clause; 0- no instance of occurrence for that syntactic purpose; dash- no occurrence of that punctuate mark for any syntactic purpose

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
Abbreviation period	38	129	6	4	0	1
List period	6	4	3	0	0	0
Decimal period	2	35	5	0	0	1
EOS period	238	254	280	16	14	47
EOS ?	6	4	8	-	1	1
EOS !	0	4	0	-	-	0

Quotations on sentence	0	4	0	-	-	-
Quotations on word	2	12	6	-	-	-
Quotations on phrase	6	10	6	-	-	-
Apostrophe-contraction	79	160	155	0	8	32
Apostrophe-plural	0	0	0	0	0	0
Apostrophe-possessive	57	63	33	7	2	3
Semicolon- list	6	2	2	-	-	-
Semicolon-main clauses	1	5	2	-	-	-
Hyphen- in a word	4	6	5	0	0	0
Hyphen- main clauses	0	1	2	0	0	0
Hyphen-main/subord.	0	0	4	0	0	0
Comma- list	42	52	7	6	2	2
Comma-main/dep.	57	102	96	5	5	3
Comma- main clauses	24	70	63	1	5	6
Comma-phrases	72	100	110	14	5	5

Table 5.15 Syntactically classified punctuation, raw

3.16 Syntactically classified punctuation- normalized per 100 for that punctuation mark

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
--	------	----	------	-------	-------	-------

Abbreviation period	12.6	30.1	2	19.1	0	.6
List period	2	.95	1	0	0	0
Decimal period	.66	8.3	1.7	0	0	1.6
EOS period	78.8	60.2	93.7	76.2	100	73.4
EOS ?	100	40	89	-	100	50
EOS !	0	80	0	-	-	0
Quotations on sentence	0	15.4	0	-	-	-
Quotations on word	22.2	46.2	42.9	-	-	-
Quotations on phrase	66.7	38.5	42.9	-	-	-
Apostrophe-contraction	57.3	70.8	80.7	0	80	91.4
Apostrophe-plural	0	0	0	0	0	0
Apostrophe-possessive	41.3	27.9	17.2	87.5	20	8.6
Semicolon- list	85.7	28.6	40	-	-	-
Semicolon-main clauses	14.3	71.4	40	-	-	-
Hyphen- in a word	2.8	2.3	3.1	0	0	0
Hyphen- main clauses	0	.39	1.2	0	0	0
Hyphen-main/subord.	0	0	2.5	0	0	0
Comma- list	21.4	15.8	2.5	23.1	11.8	14.3
Comma-main/dep.	29.1	31	34.7	19.2	29.4	21.4
Comma- main	12.3	21.3	22.7	3.9	29.4	42.9

clauses						
Comma-phrases	36.7	30.4	39.7	53.9	29.4	35.7

Table 5.16 Syntactically classified punctuation, normalized

3.17 Grapheme (unigram) profile

	Jeff	JJ	Kyle	Doc01	Doc02	Doc03
1	E 12.27%	E 11.97	E 11.8	E 10.61	E 12.25	E 13.86
2	T 9.36%	T 9.57	T 9.26	I 8.56	O 9.13	T 9.37
3	A 8.62%	A 8.87	A 8.33	N 8.4	T 9.07	N 8.73
4	O 7.26%	O 7.28	I 7.55	T 7.97	A 8.43	A 8.42
5	N 6.9	N 7.12	N 7.27	A 7.97	S 6.82	O 7.47
6	I 6.83	I 6.74	O 7.19	S 7.59	H 6.24	S 6.01
7	S 6.69	S 6.15	S 6.59	O 7.11	R 5.89	R 5.82
8	R 6.23	R 6.13	R 5.8	H 5.87	N 5.78	I 5.13
9	H 5.12	H 5.1	H 5.43	R 5.71	I 5.49	H 4.81
10	D 4.28	L 4.56	L 4.44	L 4.09	L 4.33	D 4.75
11	L 4.21	D 3.49	D 3.59	D 4.09	F 3.06	L 3.54
12	C 3.03	C 2.92	G 2.74	C 3.28	M 2.95	C 2.97
13	F 2.6	G 2.52	U 2.58	P 2.96	P 2.83	G 2.47
14	G 2.54	U 2.49	C 2.57	G 2.85	D 2.66	U 2.28
15	V 2.34	M 2.48	F 2.33	M 2.85	Y 2.48	P 2.28
16	M 2.25	P 2.2	W 2.22	F 2.26	U 2.43	Y 1.96
17	P 2.21	F 2.14	M 2.17	U 2.05	C 2.25	M 1.84
18	W 1.83	W 2.05	P 2.1	W 1.51	G 2.02	K 1.84
19	Y 1.53	Y 1.82	Y 1.79	Y 1.13	W 1.91	F 1.77
20	B 1.18	B 1.58	B 1.43	K 1.08	K 1.39	B 1.65
21	K 1.15	K 1.08	K 1.2	V .86	V 1.16	W 1.27

22	V .96	V 1.01	V 1.06	B .7	B .92	V .7
23	J .24	J .24	Z .16	J .22	X .35	X .51
24	X .18	X .23	X .15	Z .16	J .12	J .38
25	Z .13	Z .2	J .14	X .11	Q .06	Z .13
26	Q .06	Q .07	O .09	Q -	Z -	Q .06

Table 5.17 Grapheme (unigram) profile

3.18 Character n-grams, direct correlation in position

Unigrams

	Jeff	JJ	Kyle
Doc01	7	8	5
Doc02	1	2	2
Doc03	15	13	9

Table 5.18 Unigrams, direct correlation in position

Bigrams

	Jeff	JJ	Kyle
Doc01	5	7	2
Doc02	5	6	5
Doc03	6	6	5

Table 5.19 Bigrams, direct correlation in position

Trigrams

	Jeff	JJ	Kyle
Doc01	0	1	0
Doc02	1	1	3
Doc03	3	4	2

Table 5.20 Trigrams, direct correlation in position

3.19 Character bigrams and trigrams, top 10, 20, and 50 in common

Top 10 bigrams

	Jeff	JJ	Kyle
Doc01	5	5	4
Doc02	6	4	6
Doc03	7	7	7

Table 5.21 Top 10 bigrams

Top 20 bigrams

	Jeff	JJ	Kyle
Doc01	12	10	9
Doc02	11	10	10
Doc03	14	15	16

Table 5.22 Top 20 bigrams

Top 50 bigrams

	Jeff	JJ	Kyle
Doc01	31	31	32
Doc02	35	31	32
Doc03	39	36	37

Table 5.23 Top 50 bigrams

Top 10 trigrams

	Jeff	JJ	Kyle
Doc01	3	4	5
Doc02	4	1	2
Doc03	4	4	4

Table 5.24 Top 10 trigrams

Top 20 trigrams

	Jeff	JJ	Kyle
Doc01	7	6	7
Doc02	6	2	4
Doc03	8	7	7

Table 5.25 Top 20 trigrams

Top 50 trigrams

	Jeff	JJ	Kyle
Doc01	16	8	12
Doc02	12	10	12
Doc03	19	14	23

Table 5.26 Top 50 trigrams

4. Identifications per parameter

Below, the author identified for each document via each parameter is presented in tabular form. Though these results are presented in an expanded form, with individual aspects of some parameters being listed, the ultimate conclusion for each parameter was taken as a combination of the aspects to reach one most-common author conclusion. For each parameter, that result is listed in the row titled "Most common."

Notes: N/a indicates there was not enough data to eliminate even one possible author.

4.1 Average lengths

	Doc01- JJ	Doc02- Kyle	Doc03- Jeff
Avg word length	Jeff/JJ	Kyle	Kyle
Avg sentence length- words	Jeff/Kyle	JJ	Jeff/Kyle
Avg sentence length- characters	Kyle	Kyle	Jeff
Shortest sentence	N/A	N/A	N/A
Longest sentence	Kyle	Kyle	Kyle
Most common	Kyle	Kyle	Kyle

Table 5.27 Average lengths comparison

4.2 Type-token ratio

	Doc01- JJ	Doc02- Kyle	Doc03- Jeff
Ratio	Jeff	Jeff	Jeff

Table 5.28 Type-token ratio comparison

4.3 Lexical density

	Doc01- JJ	Doc02- Kyle	Doc03- Jeff
Lexical density	JJ	Kyle	Kyle

Table 5.29 Lexical density comparison

4.4 Readability- Readability Test Tool

	Doc01- JJ	Doc02- Kyle	Doc03- Jeff
Flesch Reading Ease	Jeff	JJ	Kyle
Flesch-Kincaid Grade Level	JJ	JJ	Jeff
Gunning-Fog	Jeff	JJ	Jeff
SMOG	Jeff	JJ	Jeff

Coleman Liau	JJ	JJ	JJ
ARI	Kyle	JJ	Jeff
Most common	Jeff	JJ	Jeff

Table 5.30 Readability comparison- Readability Test Tool

4.5 Readability- Online Utility

	Doc01- JJ	Doc02- Kyle	Doc03- Jeff
Flesch Reading Ease	JJ	JJ	Jeff
Flesch-Kincaid Grade Level	Kyle	JJ	Jeff
Gunning-Fog	Kyle	JJ	Jeff
SMOG	Jeff	Kyle	Kyle
Coleman Liau	JJ	JJ	Jeff
ARI	JJ	JJ	Kyle
Most common	JJ	JJ	Jeff

Table 5.31 Readability comparison- Online Utility

4.6 Parts of Speech

	Doc01- JJ	Doc02- Kyle	Doc03- Jeff
Nouns	Jeff	Kyle	Kyle
Verbs	Jeff	Kyle	Kyle
Adjectives	Jeff	JJ	JJ
Adverbs	Jeff	Kyle	JJ
Determiners	JJ	JJ	Kyle/Jeff
Existential There	JJ	Jeff/Kyle	Kyle/Jeff
Modal Auxiliaries	Jeff	Jeff	Kyle

Coordinating Conjunctions	JJ	JJ	JJ
Interjections	Kyle/JJ	Kyle/JJ	Jeff
Wh-words	Jeff	Jeff	JJ/Kyle
Prepositions or subordinating conjunctions	Kyle	Kyle	Jeff
Most common	Jeff	Kyle	Kyle

Table 5.32 Parts of speech comparison

4.7 Modals

	Doc01- JJ	Doc02- Kyle	Doc03- Jeff
would	JJ	JJ	Jeff
should	Jeff	Jeff	Jeff
shall	N/a	N/a	N/a
will	Jeff	JJ	JJ
could	JJ	JJ	Kyle
can	Jeff	Jeff	Jeff/Kyle
may	Jeff	JJ/Kyle	Jeff
must	Jeff/JJ	Jeff/JJ	Jeff/JJ
might	Kyle	Kyle	JJ
Most common	Jeff	JJ	Jeff

Table 5.33 Modal comparison

4.8 Function word profile comparison, direct correlation of function word ranking

Doc01- JJ	Doc02- Kyle	Doc03- Jeff
N/a	JJ	Kyle

Table 5.34 Function word profile comparison

4.9 Word length count total comparison, frequency

Doc01- JJ	Doc02- Kyle	Doc03- Jeff
Jeff	Kyle	JJ

Table 5.35 Word length count total comparison, frequency

4.10 Word length profile total comparison, direct correlation of ranking

Doc01- JJ	Doc02- Kyle	Doc03- Jeff
N/a	N/a	N/a

Table 5.36 Word length comparison, direct correlation of ranking

4.11 Punctuation profile comparison

	Doc01- JJ	Doc02- Kyle	Doc03- Jeff
Period	Kyle	JJ	Jeff
Comma	Kyle	JJ/Jeff	JJ/Jeff
Colon	Kyle	JJ	JJ
Semicolon	JJ	JJ	JJ
Parentheses	JJ	Jeff/Kyle	Kyle
Quotations	Jeff	Jeff	Jeff
Ampersand	JJ/Kyle	JJ/Kyle	JJ/Kyle
Plus sign	JJ	JJ	JJ
Hyphen	Jeff	JJ	Jeff
Slash	Kyle	Kyle	JJ
Ellipses	Kyle	Kyle	Jeff
Question mark	Jeff/JJ	Kyle	Kyle
Apostrophe	JJ	JJ	Kyle
Exclamation point	Kyle	Kyle	Jeff
Most common	Kyle	JJ	JJ or Jeff

Table 5.37 Punctuation profile comparison

4.12 Verb forms

	Doc01- JJ	Doc02- Kyle	Doc03- Jeff
Infinitive/ verbs	Jeff	Jeff	Kyle
Infinitive/ words	Jeff	Kyle	Kyle
Passives/ verbs	Kyle	Kyle	Kyle
Passives/ words	Jeff	Kyle	Jeff
Past part/ verbs	JJ	Kyle	Kyle
Past part/ words	Kyle	JJ	Jeff
-ing part/ verbs	Kyle	Jeff	Jeff
-ing part/ words	Jeff	Jeff	JJ
Most common	Jeff	Kyle	Kyle

Table 5.38 Verb form comparison

4.13 Syllable count comparison

	Doc01- JJ	Doc02- Kyle	Doc03- Jeff
1	JJ	Kyle	Kyle
2	Jeff	Jeff	JJ
3	Jeff	Kyle	Kyle
4	Jeff/Kyle	Jeff/Kyle	Jeff/Kyle
5	Jeff	JJ/Kyle	Jeff/Kyle
6	Jeff/JJ	Kyle	Kyle
Most common	Jeff	Kyle	Kyle

Table 5.39 Syllable count comparison

4.14 Grammatical errors

	Doc01- JJ	Doc02- Kyle	Doc03- Jeff
Sentence fragment	JJ	JJ	Jeff

Run-on sentence	Kyle	Kyle	JJ
Subject-verb mismatch	Jeff	Jeff	Jeff
Tense shift	Kyle	Jeff	Jeff
Wrong verb form	JJ	Jeff	JJ
Missing auxiliary verb	JJ/Kyle	JJ/Kyle	JJ/Kyle
Total errors	Jeff	Jeff	Kyle
Most common	JJ/Kyle	Jeff	Jeff

Table 5.40 Grammatical errors comparison

4.15 Spelling errors

There were no common spelling errors between any of the test documents and any of the author corpora.

4.16 Syntactically classified punctuation

	Doc01- JJ	Doc02- Kyle	Doc03- Jeff
Abbreviation period	Jeff	Kyle	Kyle
List period	JJ	JJ	JJ
Decimal period	Jeff	Jeff	Kyle
EOS period	Jeff	Kyle	Jeff
EOS ?	JJ	Jeff	JJ
EOS !	Jeff/Kyle	Jeff/Kyle	Jeff/Kyle
Quotations on sentence	Jeff/Kyle	Jeff/Kyle	JJ
Quotations on word	Jeff	Jeff	Jeff
Quotations on phrase	JJ	JJ	JJ

Apostrophe-contraction	Jeff	Kyle	Kyle
Apostrophe- plural	N/a	N/a	N/a
Apostrophe-possessive	Jeff	Kyle	Kyle
Semicolon- list	JJ	JJ	JJ
Semicolon- main clauses	Jeff	Jeff	Jeff
Hyphen- in a word	JJ	JJ	JJ
Hyphen- main clauses	Jeff	Jeff	Jeff
Hyphen-main/subord.	Jeff/JJ	Jeff/JJ	Jeff/JJ
Comma- list	Jeff	JJ	JJ
Comma-main/dep.	Jeff	Jeff	Jeff
Comma- main clauses	Kyle	Kyle	Kyle
Comma- phrases	Kyle	JJ	Jeff
Most common	Jeff	Jeff	JJ

Table 5.41 Syntactically-classified punctuation comparison

4.17 Grapheme profile comparison

The grapheme profile comparison is reflected in the direct correlation of position comparison of unigrams.

4.18 Character n-gram comparison, direct correlation of position

	Doc01- JJ	Doc02- Kyle	Doc03- Jeff
Unigrams	JJ	JJ/Kyle	Jeff
Bigrams	JJ	JJ	Jeff/JJ
Trigrams	JJ	Kyle	JJ

Most common	JJ	JJ/Kyle	Jeff/JJ
-------------	----	---------	---------

Table 5.42 Character n-gram comparison

4.19 Character bigram and trigram comparisons, top 10, 20, and 50 in common

	Doc01- JJ	Doc02- Kyle	Doc03- Jeff
Top 10 bigrams	Jeff/JJ	Jeff/Kyle	N/a
Top 20 bigrams	Jeff	Jeff	Kyle
Top 50 bigrams	Kyle	Jeff	Jeff
Most common	Jeff	Jeff	Jeff/Kyle

Table 5.43 Top bigram comparisons

	Doc01- JJ	Doc02- Kyle	Doc03- Jeff
Top 10 trigrams	Kyle	Jeff	N/a
Top 20 trigrams	Jeff/Kyle	Jeff	Jeff
Top 50 trigrams	Jeff	Jeff/Kyle	Kyle
Most common	Jeff/Kyle	Jeff	Jeff/Kyle

Table 5.44 Top trigram comparisons

5. Overall identification success per parameter

Below, the overall total results for each parameter are presented in tabular form, with the percentage of correctly-identified authors and a statement on whether the percentage is high enough to be better than chance (50%). Results where two possible authors were identified were counted as incorrect, regardless of whether one of the two authors was the correct one. While in these cases the parameter's ability to potentially narrow author pool is worth further investigation, based on the design of this study, their inability to narrow the field to a single author is a failure of the parameter to identify the author correctly.

	Percentage correct	Better than chance?
Average lengths	33%	No
Type-token ratio	33%	No
Lexical density	66%	Yes
Readability- Readability Test Tool	33%	No
Readability- Online Utility	66%	Yes
Modals	33%	No
Function word profiles	0%	No
Word length frequency	33%	No
Word length profile	0%	No
Punctuation profile	0%	No
Verb forms	33%	No
Syllable count	33%	No
Grammatical errors	33%	No
Spelling errors	0%	No
Syntactically-classified punctuation	0%	No
N-gram frequency position	33%	No
Top bigrams	0%	No
Top trigrams	0%	No

Table 5.45 Accuracy

Chapter 6

What Does It All Mean, and What Is the Next Step?

1. Summary and discussion of results

Via this study, I set out to examine a variety of parameters, most of which have been utilized in previous studies, for their viability as determiners of authorship in a forensically-feasible setting and on a dataset representative of an under-researched register. Blog posts have not often been the subject of authorship work, and blog posts present a forensically-feasible register, as a circumstance where law enforcement agencies are attempting to determine the true author of an incendiary or threatening blog post, for example, is certainly conceivable. The majority of authorship studies have examined documents along only one identification parameter or along only a small, related group of parameters, such as several different types of n-grams. The few studies that have been conducted using a large body of parameters have not utilized internet-based text, including blog posts, and this is a medium rich with forensically-feasible data which offers some very different characteristics and thus unique challenges where concerns authorship determination. This study was designed to help fill these gaps in the research in a field where a deep body of research is necessary to move the forensic aspect- and, in particular, the aspect of court acceptability- forward.

I presented an overview of the history and characteristics of this register, blogs and the posts within, and then carried out a register analysis based on Biber and Conrad's (2001) methodology on the specific blog I intended to use for my authorship study. I then developed corpora for each of three chosen blog authors from the Detroit Red Wings hockey blog *Winging It in Motown* and examined those

corpora per the list of chosen parameters. I also chose a test document from each author, not already included in that author's corpus, and coded the documents with numbers. Once the corpora had been tested against the chosen parameters, I examined each document for the same parameters and then carried out a comparison between each author corpus and each test document along each parameter to determine which corpus most closely matched each test document along each parameter. These results were recorded and then examined against a key identifying the author of each test document for accuracy of identification.

The results of the study indicate that most of the parameters did not successfully distinguish among the three possible authors in this data set. The requirement for the parameters to have been successful in distinguishing authors was that they correctly identified the author of at least two of the three test documents, which would place the likelihood of correct identification at a higher percentage than mere chance. However, only two of the parameters, lexical density and readability via Online Utility's readability utility, achieved this level of accuracy. Furthermore, one of the parameters that accurately identified two documents' authors, the Readability via Online Utility parameter, is suspect, because this parameter was tested via two separate utilities designed to review readability, Online Utility and Readability Test Tool, and while Online Utility succeeded in identifying two texts' authors, Readability Test Tool, using the same formulas but likely different definitions for counts of words and word types, failed to achieve the same level of success. This suggests that the parameter still may have simply correctly identified two authors by chance. It further suggests that the parameter is likely still not reliable or replicable, thus requiring further studies before suggesting that it may be a viable option for authorship work.

The success of lexical density as a parameter was somewhat unexpected. This parameter essentially measures the ratio of lexical words within the context of all words in a document. The possibility exists that the habit to be more or less detailed in one's language use, which could be reflected by a higher or lower portion of the total words being lexical, may be idiosyncratic. Thus, a high lexical density ranking would suggest that the individual includes more detail, reflected by lexical words, in a habitual manner, which would be a characteristic that would reflect author idiolect. However, this parameter is not without problems, which tempered my expectations for its performance before the study began. Namely, as Biber et al. (1999) have shown, certain registers of text showcase linguistic usages that would likely coincide with a high lexical density measurement such as particularly high ratios of nouns, verbs, adjectives, and adverbs- all lexical categories- to determiners, prepositions, modals, and other categories considered to be function words. Registers with a highly informative purpose, such as academic texts and news, do tend to be more densely packed, as the creators of those texts attempt to fit a significant amount of information into a small space. As established in chapter 3, WiiM is, overall, highly informative in nature, but its purpose is hybrid, and authors do also express some opinion via their posts. It is thus not out of the realm of possibility that specific purposes of each post may drive the difference. Perhaps JJ writes more informative posts while Kyle authors many opinion pieces, or perhaps this is not true overall but it is true within the time period from which the corpus posts and the test documents were selected. In this case, lexical density could be driven by something besides author idiolect, something related to the subtopic or subregister of the post. Though related parameters, such as type-token ratio and parts of speech counts, have been examined in multiple research studies, little work focusing on lexical density has been carried out, so further research to attempt to replicate this result, especially

while controlling for potential subtopic and subregister effects, must be carried out before this parameter can be viewed as a potential option for authorship work.

While my expectations for most parameters tested were low, I did not anticipate such a low success rate for the entire study. In particular, I had high expectations for syntactically-classified punctuation. I also expected to see potentially strong results from verb forms. I believed that these two parameters, which should offer some reflection of deeper syntactic structure, had the most potential to differentiate. This idea was largely driven by Chaski's (2001) assertions that deep syntactic structures are largely created or chosen subconsciously and that parameters that reflect syntactic use are rooted in linguistic theory, as well as that these types of parameters tend to provide more data to work with in short texts than other parameters often used in literary analysis work with large documents. I retain my belief that this type of approach is more likely to capture author idiolect and thus to provide an accurate and reliable means of determining or authenticating authorship, particularly when working with short documents from registers likely to show up in forensic contexts. These types of parameters need further, broader study to obtain a clearer picture of their abilities and limitations within the context of authorship work.

2. Challenges and Limitations of the study

While all efforts were made to ensure this study was as robust as possible given the scope available, there are limitations in place which warrant discussion. The lack of ability to develop automated systems restricts the amount of data which could be processed and examined in the time period given. This limitation restricted the possible size of the author corpora, which was kept manageable by utilizing only

10 blog posts from each author, as well as the number of authors whose texts could be examined. WiiM is a highly prolific blog with multiple contributors, and there are more than three members who blog frequently. This problem also restricted the number of test documents that could be handled for the examination. Ideally, a study such as this would be run against numerous text documents from each author. That the quantity of data that could be handled for this study was restricted is not ideal in the testing phase, but it is a much more accurate reflection of a real-life forensic situation, when officials often have little data of either type, known or test, to work with. Testing methods against forensically-realistic datasets also offers important contributions. The issue of the time and man-power constraints presented by manual examination of the texts also limited the number and nature of possible parameters to be examined. Though many possible parameters were featured in this study, numerous further options exist that could also be explored. This presented the most significant source of frustration in the study, as ideally as many parameters as can be imagined would be tested in order to find parameters that work reliably as well as to narrow down what driving forces may be behind the success of certain parameters to open doors to new possibilities.

Demographics also presented a challenge during this study. The three main bloggers who were featured in this study are all male. They fall into a close age range, are all college-educated, and are all white. While they hail from different locales, they all appear to exist in similar environments. This presents a positive aspect as well, namely that distinguishing between authors with very similar demographic make-ups can present an extra challenge and is also very much a forensically-realistic situation, as having possible suspects who are very similar is a

common circumstance. Thus, this limitation comes with something of a silver lining as well.

3. Future directions

There are numerous approaches one can take to further the research presented in this study and related research in the field. As the parameters examined presented very little success in this study, a promising next step would be running the data through Chaski's software to see if her proprietary method is more successful at accurately identifying the author of each test document than this study's parameters were. Her software relies primarily on syntax-driven parameters and she has reported high levels of success with her methods relative to other studied parameters and other methods of utilizing those parameters (Chaski, 2007). However, as her methods are proprietary and her software is in the process of becoming patented, replicating her studies is more complex, as is reporting how they function and dissecting why they might work well, and doing so was beyond the scope of this dissertation.

Other possible future directions involve continuing to search for other possible parameters and studying these same datasets against them, as well as expanding the datasets and re-testing them to attempt to discover whether the parameters may have been more successful with larger author corpora and/or longer test documents. Including more bloggers in the study is another possible angle for future research, in order to enlarge the pool of possible authors. As WiiM is an extensive and extremely active blog with multiple authors, it is possible that lower-level bloggers may also provide enough data to make a larger study examining more authors at once possible. Bringing into this study the ability to code linguistic utilities to automate

examination of the parameters could potentially also offer altered results, as such utilities would remove the human element and allow for more complex comparisons.

REFERENCES

- Bar-Ilan, J. (2005). Information hub blogs. *Journal of Information Science*, 31(4), 297-307.
- Biber, D., & Conrad, S (eds). (2001). *Variation in English: Multi-Dimensional Studies*. Harlow: Pearson Education Limited.
- Biber, D., & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Essex, England: Pearson Education.
- Blood, R. (2000). Weblogs: A history and perspective. *Rebecca's Pocket*. Retrieved from http://www.rebeccablood.net/essays/weblog_history.html
- Blood, R. (2002a). Introduction. In Rodzvilla, J. (Ed.), *We've Got Blog: How Weblogs Are Changing Our Culture* (pp. ix-xii). Cambridge, MA: Perseus Publishing.
- Blood, R. (2002b). *The Weblog Handbook: Practical Advice on Creating and Maintaining Your Blog*. Cambridge, MA: Perseus Publishing.
- Blood, R. (2003). Weblogs and journalism in the age of participatory media. *Rebecca's Pocket*. Retrieved from http://www.rebeccablood.net/essays/weblogs_journalism.html
- Blood, R. (2004). How blogging software reshapes the online community. *Communications of the ACM*, 47(12), 53-55.
- Brala, M. (2008). Language, policy and identity; Perceptions of and expectations for (non)Anglicized language on the web. The case of Croatian blogs. *Bulletin Suisse de Linguistique Appliquee*, 87, 73-94.
- Carvin, A. (2007, December 24). The evolution of the blog. *NPR*. Retrieved from <http://www.npr.org/templates/story/story.php?storyId=17421022>
- Chaski, C. E. (1999). Linguistic authentication and reliability. In *National Conference on Science and the Law Proceedings, NIJ Research Forum*. Paper presented at the National Conference on Science and the Law, San Diego, California, 15-16 April (pp. 97-140).
- Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8(1), 1-65.
- Chaski, C. E. (2007). The keyboard dilemma and authorship identification. In P. Craiger & S. Sheno (Eds.), *Advances in Digital Forensics III* (pp. 135-146). New York, New York: Springer.
- Chaski, C. E. (2013). Best practices and admissibility of forensic author identification. *Brooklyn Journal of Law and Policy*, 21(2), 333-376.

- Copeland, H. (2004). Blog reader survey May 17-19, 2004. Retrieved from http://www.blogads.com/survey/blog_reader_survey.html
- Cox, T. (2014). *The Language of Winging It in Motown: A Register Analysis of a Sports Blog*. Unpublished manuscript.
- Crystal, D. (2009). *Language and the Internet*. Cambridge: Cambridge University Press.
- Damereau, F. J. (1975). The use of function word frequencies as indicators of style. *Computers and the Humanities*, 9, 271-280.
- Dardick, G. S., La Roche, C. R., & Flanigan, M. A. (2007). Blogs: Anti-forensics and counter anti-forensics. *Proceedings of the 5th Australian Digital Forensics Conference*. Perth, Western Australia: Edith Cowen University.
- De Moor, A., & Efimova, L. (2004). An argumentation analysis of weblog conversations. *Proceedings of the 9th International Working Conference on the Language-Action Perspective on Communication Modelling*. New Brunswick, New Jersey: Rutgers.
- De Morgan, S. (1882). *Memoir of Augustus de Morgan by His Wife Sophia Elizabeth de Morgan with Selections from His Letters*. London: Longmans, Green, and Co.
- Eddy, H. T. (1887). The characteristic curves of composition. *Science*, 9(216), 297.
- Foster, D. (2002). *Author Unknown*. London: Macmillan.
- Garden, M. (2012). Defining blog: A fool's errand or a necessary undertaking. *Journalism*, 13(4), 483-499.
- Gif. (n.d.). In *Merriam-Webster online*. Retrieved from <https://www.merriam-webster.com/dictionary/GIF>
- Grant, T. (2013). TXT 4N6: Method, consistency, and distinctiveness in the analysis of SMS text messages. *Journal of Law and Policy*, 21(2), 467-494.
- Grieve, J. W. (2005). *Quantitative Authorship Attribution: A History and an Evaluation of Techniques* (Master Thesis). Simon Fraser University, Burnaby, British Columbia.
- Herring, S. C., Scheidt, L. A., Wright, E., & Bonus, S. (2005). Weblogs as a bridging genre. *Information Technology & People*, 18(2), 142-171
- Holmes, D. I. (1992). A stylometric analysis of Mormon scripture and related texts. *Journal of the Royal Statistical Society*, 155(1), 91-120.
- Hood, E. (2013). The #1 small business marketing idea [Infographic]. Retrieved from <http://blog.ignitespot.com/blog/small-business-marketing-idea>

- Juola, P. (2012). Large-scale experiments in authorship attribution. *English Studies*, 93(3), 275-283.
- Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1), 9-26.
- Luyckx, K., & Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, 26(1), 35-55.
- Mascol, C. (1888). Curves of Pauline and Pseudo-Pauline Style I. *Unitarian Review*, 30, 452-460.
- McMenamin, G. R. (2002). *Forensic Linguistics: Advances in Forensic Stylistics* (Kindle version). Boca Raton: CRC Press LLC.
- Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, 9(214), 237-249.
- Morton, A. Q. (1978). *Literary Detection: How to Prove Authorship and Fraud in Literature and Documents*. New York: Scribners.
- Mosteller, F., & Wallace, D. (1964). *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley.
- Myers, G. (2010). *The Discourse of Blogs and Wikis*. London, UK: Continuum.
- Nielsen.com. (2012). Buzz in the blogosphere: Millions more bloggers and blog readers. Retrieved from <http://www.nielsen.com/us/en/insights/news/2012/buzz-in-the-blogosphere-millions-more-bloggers-and-blog-readers.html>
- Nini, A. (2014). *Multi-dimensional Analysis Tagger 1.2- Manual*. Retrieved from: <http://sites.google.com/site/multidimensionaltagger>
- Nowson, S., Oberlander, J., & Gill, A. J. (2005). Weblogs, genres, and individual differences. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, (pp. 1666–1671). Hillsdale, NJ: Lawrence Erlbaum Associates.
- O'Donnell, B. (1966). Stephen Crane's *The O'Ruddy*: A problem in authorship discrimination. *The Computer and Literary Style*. Kent State University Press: 107-115.
- Pingdom.com. (2013). Blog readership demographics- investigating the world's top blogs. Retrieved from <http://royal.pingdom.com/2013/03/01/blog-readership-demographics-2013/>

- Salen, T. (2007). *Weblogs and Blogging: Constructivist Pedagogy and Active Learning in Higher Education* (Doctoral Dissertation). University of Bergen, Norway.
- Santorini, B. (1990). *Part-Of-Speech Tagging Guidelines for the Penn Treebank Project (3rd revision, 2nd printing)*. Department of Linguistics, University of Pennsylvania.
- Schmidt, J. (2007). Blogging practices: An analytical framework. *Journal of Computer-Mediated Communication*, 12, 1409-1427.
- Stamatatos, E. (2013). On the robustness of authorship attribution based on character N-gram features. *Journal of Law and Policy*, 21(2), 421-439.
- Statista.com. (2017a). Cumulative total of Tumblr blogs from May 2011 to April 2017 (in millions). Retrieved from <https://www.statista.com/statistics/256235/total-cumulative-number-of-tumblr-blogs/>
- Statista.com. (2017b). Number of blogs worldwide from 2006 to 2011 (in millions). Retrieved from <https://www.statista.com/statistics/278527/number-of-blogs-worldwide/>
- Statista.com. (2017c). Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2017 (in millions). Retrieved from <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>
- Stone, B. (2004). *Who Let the Blogs Out?: A Hyperconnected Peek at the World of Weblogs*. New York, NY: St. Martin's Griffin.
- Svartik, J. (1968). *The Evans Statement*. Gothenburg: University Gothenburg.
- Technorati.com. (2010). State of the blogosphere 2010. Retrieved from <http://technorati.com/state-of-the-blogosphere-2010/>
- Walker Rettberg, J. (2014). *Blogging*. Cambridge, UK: Polity Press.
- Weil, D. (2006). *The Corporate Blogging Book: Absolutely Everything You Need to Know to Get It Right*. New York, NY: Portfolio.
- Woolls, D. (2012). Detecting Plagiarism. In P.M. Tiersma & L.M. Solan (Eds.), *The Oxford Handbook of Language and Law* (Kindle version). Oxford: Oxford University Press.
- Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. *Biometrics*, 31, 356-361.

APPENDIX I
TEST DOCUMENTS

TEST DOCUMENT 1

GRAND RAPIDS, Mich. – The **Detroit Red Wings** on Wednesday reassigned forward **Givani Smith** (*jih-VAH-nee SMITH*) to the Grand Rapids Griffins from the Ontario Hockey League's Guelph Storm and defenseman **Filip Hronek** (*FIHL-ihp H'RAWN-ehk*) to the American Hockey League club from the OHL's Saginaw Spirit.

Smith, 19, played in 64 games with Guelph this season and ranked among the team's leaders with 44 points (3rd), 26 goals (2nd), seven power play goals (T2nd), three game-winning goals (T1st) and 214 shots (1st). The 6-foot-2, 209-pound winger logged 139 penalty minutes to lead the league for the second consecutive season.

Detroit's second choice (46th overall) in the 2016 **NHL Entry Draft**, Smith has accumulated 101 points (56-45—101) in 159 games since joining Guelph midway through the 2014-15 campaign. In his first full season with the Storm in 2015-16, he finished second on the club with 42 points (23-19—42).

A native of Thornhill, Ontario, Smith split his initial OHL season in 2014-15 between Barrie and Guelph. He chipped in four assists and 20 PIM in 31 games with the Colts before totaling 15 points (7-8—15) and 56 PIM in 30 games with the Storm to conclude the season. He added five points (2-3—5) in nine playoff contests.

In his first season in North America, Hronek, 19, skated in 59 games with Saginaw this year and tied for fourth among OHL defenseman in scoring (14-47—61). Detroit's third choice (53rd overall) in the 2016 NHL Entry Draft, Hronek placed among the team's leaders with 47 assists (1st), 21 power play points (1st), 14 goals (4th) and 235 shots (2nd) and was named Saginaw's Most Valuable Player.

A native of Hradec Kralove, Czech Republic, Hronek has represented his home country in international competition since the 2014-15 season, and most recently, notched four points (2-2—4) and 10 PIM in five games at the 2017 World Junior Championship.

Prior to his North American debut, the 6-foot, 170-pound blueliner played in the Czech Republic from 2013-16. Skating with the Hradec Kralove U18 and junior team as well as Litomerice, Hronek appeared in 124 games totaling 76 points (23-53—76) and 220 PIM. He debuted professionally with Hradec Kralove in the Czech Extraliga in 2014-15 and appeared in 41 games across two seasons with the club, picking up four assists and 24 PIM.

The Central Division-leading Griffins host Rockford on Friday at 7 p.m.

Single-game tickets are currently on sale. Fans can secure their **full-season**, **select-season** or **group** ticket packages by calling (616) 774-4585 ext. 2 or visit griffinshockey.com for more information.

TEST DOCUMENT 2

The SHL playoffs are underway, and Red Wings prospect Axel Holmstrom has come out of the gate with his offense surging against Frolunda. As many of you know, another Red Wings prospect, Christoffer Ehn, plays with Frolunda, so both youngsters have had the daunting task of going head-to-head — something not many future teammates get to do over in Europe.

Holmstrom has put up five points (3-2-5) in three games so far — impressive for the 20-year-old, who is now considered an SHL veteran. As it stands now, Holmstrom has put up 29 points (12-17-29) in his 28 playoff games in the SHL. It was only a couple of seasons ago did the former 7th-rounder make headlines as he scored at a historic rate in the 2014-2015 playoffs.

Patrik Bexell, a writer over at EOTP and Euro hockey guru caught up with Holmstrom and his coach to talk about being back in the playoffs, expectations, and other stuff. Listen to the raw audio:

While Holmstrom seems to have confidence in his game and continues to flourish for his team on a big stage, Skellefteå AIK head coach Stefan Clockare spoke about Holmstrom's talent and how he's responded after recovering from injuries:

Now, as many of you are probably wondering, when will Holmstrom make the move to North America? It's likely he may come over after the SHL playoffs to join the Griffins as they make a push for the playoffs, but that remains to be seen. Playoff time isn't exactly the best time to ask players or coaches about their plans to leave Europe for another league, but take note of this — Holmstrom signed a three-year contract with Skellefteå as a 17-year-old back in 2014. The Red Wings have expressed that they are comfortable with letting him develop at a high level in Sweden until they feel he's ready to make the move. I'd say there is a good chance he spends another year or so in the SHL.

Either way, Holmstrom's style of play will translate over to North American hockey quite well. He's a strong two-way center who plays hard on the puck and goes to the net. While I don't see him as a top-six guy, there's plenty of room left for him to develop into a very special player.

Everyone give Patrik a [**follow on Twitter**](#) to keep up to date with happenings in European hockey.

TEST DOCUMENT 3

Hey there... ready for the latest edition of Pro or No?

Today's contestant is non other than Brad Richards.

Mr. Richards came to the Red Wings last offseason on a 1-year deal with dreams of winning another **Stanley Cup**. He was supposed to be that key veteran piece that could play 2nd line center and allow **Henrik Zetterberg** and **Pavel Datsyuk** to play together. But much like every other player that's recently been put into that position... he failed.

Brad Richards

#17 / Center / **Detroit Red Wings**

Height: 6-0

Weight: 199

Born: May 2, 1980

The Pro

Let's be honest here... there aren't many pros. Quite frankly, I'm not even 100% sure what to write. Brad Richards seems like a nice guy. He's a vet that brings experience and leadership to the dressing room. He stays out of the box. Only 8 PIM in 68 games. Richards clearly doesn't put his team on the disadvantage. The Wings also had the puck more often than not when he was on the ice. His **CF% in all situations** was 54.7%. Richards has a nice smile too. Oh! There was that one time he scored the game winning goal in the Stadium Series game in Colorado and sent me, Kyle, JJ and Graham home happy. Thanks for that, Brad.

The No

He's old.

The Red Wings don't need more old players these days. There are enough vets in the room to fill that role. A 36-year-old on the decline is not in the cards for this team. Leave the old vet leadership role to Zetterberg and Niklas Kronwall. Let guys like **Justin Abdelkader** and **Danny DeKeyser** step up and fill any leadership void there might be. The Wings are team desperate for an infusion of youth. No more signings that are expected to bridge the gap from the old guard to the new. We're past that. If this truly is a summer of change, allowing Richards to walk and not signing anyone similar is what needs to be done.

He doesn't score anymore.

Richards' 28 points this past season was the lowest output he's had since... forever. Even in the lockout season 3 years ago he put up 34 points in 46 games. I understand he wasn't playing with **Patrick Kane** anymore... but come on, the Wings players aren't THAT bad (I think). Richards points per game have declined steadily over the last 5 seasons. Not totally surprising since he's getting up there in age, but there's zero reason to think that trend won't continue and his totals will take another dip next year.

He's not a 2nd line center.

As previously mentioned, the Red Wings initially envisioned Richards as a 2C between **Tomas Tatar** and Gus Nyquist. Just like **Valtteri Filppula** and **Stephen Weiss** before him, Richards couldn't fill that role. He didn't really seem to mesh well with anyone on the roster. He spent most the season playing on Datsyuk's wing along with Darren Helm. Basically an entire line of guys who probably won't be on this team anymore.

He's not worth the money.

I think it's pretty clear if Richards were to come back to the Wings, he wouldn't be getting another deal worth \$3 million. The Wings didn't win a playoff round, so Richards bonuses didn't kick in. I guess that's a plus to being one-and-done again. I can't imagine Richards taking a huge pay cut. He's not worth \$3 million, but he's not going to come back for \$500k. Don't pay the old guy. Just don't do it.

The Verdict?

I think this one is pretty clear, but it's up to you to vote and decide.