GÖTEBORGS UNIVERSITET
INST FÖR SPRÅK OCH LITTERATURER

ENGLISH

Formality and contextuality in blogs

A linguistic analysis

Sarah Eldursi

MA thesis
Spring 2013

Supervisor:
Joe Trotta
Examiner:
Mats Mobärg

Title: *Formality and contextuality in blogs: A linguistic analysis*

Author: Sarah Eldursi
Supervisor: Joe Trotta

Abstract:

The aim of this study is to investigate formality and contextuality in weblogs. It investigates the main formality and contextuality features principally focusing on the F-score proposed by Heylighen and Dewale (1999). The primary material comes from the online blog directory *Technorati* and from *Blogeries.com*. The method was a corpus analysis using the concordance software *Wordsmith* in order to carry out a linguistic analysis of the data. The main findings confirm previous findings that blogs are mainly separated into thematic and Personal blogs. In addition, the more focused the author on imparting information, the higher the F-score and consequently the lower the contextuality of the text. Authors who focused on their personal lives produced more contextual texts with lower F-scores. The study suggests that whilst the F-score is a good indicator of formality, it should not be considered an absolute indication of formality in the traditional sense, rather it should be seen as an indication of a text's contextuality and be used as a basis from which further investigation can be developed.

Keywords: weblogs, formality, F-score, contextuality, pronouns, contractions, hedges, emphatics.

Acknowledgements

I would like to express my deepest appreciation and gratitude to my supervisor Dr. Joseph Trotta for his advice and guidance throughout writing this thesis.

I would also like to thank my beloved husband Muath for his support and encouragement for which I am greatly indebted. I would also like to thank my beautiful daughter Dalia for her love and inspiration. I similarly extend my sincere gratitude to my parents whose love and support have enabled me to become the person that I am today.

This thesis is dedicated to my husband.

Contents Page

1. Introduction 6
2. Aim 7
3. Background 8
3.1. Basic concepts 8
3.1.1. Blogs 8
3.2. Formality and contextuality 13
3.2.1. Formal language 13
3.2.2. Contextuality and Formality 14
3.2.3. The F-score 17
3.3. Previous research 19
3.3.1.Herring and Paolillo (2006) 19
3.3.2. Nowson et al. (2005) 20
3.3.3. Teddiman (2009) 21
3.3.3. Daems et al. (2013) 23
3.3.4. Grieve et al. (2010) 24
3.3.5. Biber (1988) 26
4. Material 30
5. Method 32
6. Results 33
6.1. F-score results 33
6.2. Word classes for each blog 35
6.3. Type-token ratio 38
6.4. Key words 39
6.5. Pronouns 42
6.5.1. Personal pronouns 42
6.5.2. Demonstrative pronouns 45
6.5.3. Possessive pronouns 46
6.6. Contractions 47
6.7. Hedges 49
6.7.1. Hedges listed in Knight et al. (2013) 49
6.7.2.Hedges listed in Biber (1988) 50
6.8. Downtoners 51
6.9. Amplifiers 52
6.10. Emphatics 53
6.11. Discourse particles 54
6.12. Private verbs 55
6.13. Modal and semi modal verbs 56
6.14. Expletives 58
6.15. Word and sentence length 59
6.16. F-scores for individual blogs 60
6.16.1 Political blogs 60
6.16.2 Finance blogs 61
6.16.3. Art blogs 62
6.16.4. Family blogs 62
6.16.5. Food blogs 63
6.16.6. Sport blogs 65
6.16.7. Celebrity blogs 66

6.16.8 Personal blogs 66
7. Discussion 67
References 72
Appendix A 75
Appendix B 75
Appendix C 76
Appendix D 77

1. Introduction

A number of studies have investigated the formality of language and the differences between registers and genres and this has continued with the rise and popularity of Computer Mediated Communication (CMC). In the early days of the Internet, many linguists examined the nature of CMC in terms of it being a hybrid between speech and writing. Crystal (2001) initially termed CMC 'Netspeak' and characterized it with a number of features including abbreviated forms, emoticons, etc. Although he has since abandoned this term in favor of *Internet linguistics*, which refers to "the scientific study of all manifestations of language in the electronic medium" (Crystal 2011:11), it has now been established that online communication contains features from both speech and writing with some varieties of CMC more speech-like than others (Crystal 200l, Herring 2011). Herring (2011) further provides an overview on research on CMC and conversation and cites scholars as early as Horowitz and Samuels (1987) who described CMC as "Speech writ down", Maynor (1994) who described it as "written speech" and Sack (2000) who characterized Usenet newsgroups and forums as "very larger scale conversations". Herring (2011:7) concludes that scholarly research on CMC has shown that "CMC is 'talk' and 'conversation'". Herring does however acknowledge that some modes of CMC may not be generally conversational but that users may use them in that way such as weblogs, *Wikis* and *YouTube* (Herring 2011:5).

The development of smart phones and tablets as well as 'Web 2.0' user-generated content has provided linguists with even more language styles to examine. It has also emphasized the problem with the term CMC, which some scholars (e.g. Baron 2008, Crystal 2011) state can no longer be accurately used to refer to the language used, as computers are no longer the only medium through which communication occurs. Baron further states that the fast developments in technology, which have resulted in devices such as BlackBerrys and smartphones, have deemed the term CMC somewhat inappropriate since these devices are not computers and suggests the term "*electronically mediated communication*" instead (Baron 2008:12, italics mine). Baron is not alone in suggesting a new term to use rather the CMC. Crystal 2011, as stated above, proposed the term *Internet Linguistics*.

Along with developments in new media, which allow users to communicate using a range of programs, we also have the concept of Web 2.0 developed in 2004 for a conference of the same name to refer to websites that use technology that is developed and more advanced than static webpages and includes social networking sites such as *Facebook*, *Twitter*, *Flickr*, *Instagram* and blogs. The term Web 2.0 is closely associated with Tim O'Reilly who states that the name came into being after a brainstorming session with a

number of colleagues (O'Reilly 2007). In light of these developments, there has been a move to use new terms such as the term *New Media* instead of the term CMC. Other terms that are used include *e-language* and *digital discourse*. Nevertheless, it must be noted that the term CMC remains popular and is still widely used in current linguistic research (e.g. Herring 2011). This discussion has provided a brief overview of the history of CMC and the discussion will now turn to blogs and the organization of this study.

Weblogs, which have seen an intense rise in recent years, are one of the areas on which linguistic research has focused. Word press.com, one of the world's most popular blog publishing tools, estimated that, in 2012, more than 399 million people viewed 3.7 billion blogs every month (Word press [online]). This dramatic proliferation of blogs in recent years provides a valuable resource of linguistic data and this study aims to investigate the level of formality and contextuality in blogs according to blog topic.

This study is structured according to the following organizational principles: First the focus and aim of the study are introduced. Secondly, this paper explains the basic terms and concepts in the study of CMC and blogs. The study then provides an overview of some of the most influential research in this field followed by an explanation of the main methodological approaches used in this study and finally a discussion of results ensues.

2. Aim

This study aims to investigate the linguistic differences between blogs based on blog topic/ genre and addresses the following:

1. The degree of linguistic formality and contextuality in blogs
2. The question of how topic affects formality
3. Compare frequencies of some formal features to the Brown Family of corpora

A linguistic analysis of blogs is of significant value given their immense popularity and would contribute to an understanding of this type of language of variety. Moreover, examining linguistic variations of blogs could have implications for blog searches and assist blog search engines, which could be enhanced by linguistic based searches. In addition, as indicated by Daems et al. (2013) little is known about the linguistic properties of blogs and more research is needed in this area

3. Background

This section outlines the structure of blogs and provides an explanation of the basic terms and concepts essential to this thesis.

3.1. Basic concepts

3.1.1. Blogs

Blogs are generally defined as frequently updated webpages arranged in reverse chronological order so that the most recent entry appears first. According to Blood (2000), an authority on weblogs and the author of author of the *Weblog Hand Book,* the word *weblog* was coined by the American blogger Jorn Barger in 1997. The term *blog* came into being after another blogger, Peter Merholz, decided to use word play on the word *weblog* and called it 'wee blog' which eventually became the word blog. [1] The real breakthrough for blogs came with the development of blog publishing tools in 1999, which enabled those with no programming knowledge to easily and freely publish online, which in turn opened up avenues that where previously limited to the tech savvy.

Herman et al. (2005:45) state that "a weblog, or blog, is a frequently updated website consisting of dated entries arranged in reverse chronological order so that the most recent post appears first". In terms of the main features of blogs, Baron (2008:11) states that all blogs share four basic features which are that they are predominantly text based, they are in reverse chronological order, frequently updated and they link to other sites.

The term blog is now more frequently used instead of weblog and in 2004 the Merriam-Webster dictionary identified the term *blog* as the word of the year. The Merriam-Webster dictionary defines a *blog* as "a website that contains an online personal journal with reflections, comments and often hyperlinks provided by the writer" (Blog 2013, Merriam Webster [online]). This definition emphasizes the common perception that blogs focus on presenting more personal information than impersonal, objective content. However, this is not always the case, and to further understand the characteristics of blogs and explore the distinctions between the different blog types an overview of the main investigations into blog types is presented.

Blood (2000) points to three different types of blogs: *Filters* which are blogs about issues external to the bloggers personal life, *Personal journals* which are blogs pertaining to

[1] It is widely accepted that the first weblog was the first website created by Tim Berners-Lee (1991) who created the World Wide Web at CERN where he provided links to other sites as they came online.

the blogger's personal life and *Notebooks* which are long essays and may be about issues external to the blogger's life or issues that concern the blogger's life. Blood (2000) stated that although filter blogs were the first type of blog, Personal blogs are the most common. Blood also claims that blogs are native to the web rather than being carried over from other offline genres, which is a slightly different view from Herring et al.

Herring et al. (2004) developed Blood's 2000 classification of blogs for their study on blog genres. However, they found that *Notebook* blogs were very rare, in addition to the difficulty of using length to classify them as a separate genre. Thus Herring et al. (2004) examined their corpus in terms of the purpose of the blogs, which is a key criterion for defining a genre. The blog types they found were *Filter blogs*, *Personal journals*, *Knowledge blogs* (k-log) and *Mixed purpose blogs*. Their results showed that the two main blog types represented in their random corpus were first the personal journal blog, which accounted for 70.4% of the blogs, and second filter blogs, which were much less at 12.6%. The other blog types such as K-logs featured at 3.2%, mixed blogs featured at 9.5% and other blogs featured at 4.5%. Herring et al. (2004) show that Personal blogs in particular are the most frequent type of blog online followed by filter blogs. Thus most blog classifications often focus on these two blog types. Greive et al. (2010), who carried out a multidimensional analysis of weblogs also concur that the main distinction that is made with regards to blogs is between personal versus thematic blogs.

Nowson (2006:32) recognizes three main types of blogs: *News*, *Commentary* and *Journal*. News blogs collect news on different topics and are often updated a number of times a day. Commentary blogs also focus on external material, but are not under the time constraints as the news blogs are and contain more personal input. Journal weblogs are "simply online diaries" (Nowson 2006:35) which focus on the internal workings of the blog writer. However, Herring et al.'s classification of blog types into filters, knowledge logs and personal journals remains one of the most influential blog classifications, and upon closer inspection, it becomes apparent that Herring et al.'s classification is similar to the three main blog categories identified by Nowson (2006).

This type of discussion of the different blog classifications aids our understanding of the blog genre. For the establishment of genre to be relevant, it has to be recognized that weblogs are not just texts but also include other elements such as images, videos, adverts and links. The extent to which blogs incorporate links varies with some blogs consisting mainly of links whilst others contain very few. Links, however, remain important to blogs and are even termed "the currency of the blogosphere"(Myers 2010:24). The term *blogosphere* itself was

coined in 1999 by Brad L. Graham to refer to the blogging community. The linking of *YouTube* videos in particular is of particular significance to blogs as it can increase blog views. In addition to linking to *YouTube*, many bloggers also link to their own Personal page, *Flickr* and other personal photos. Myers also found that many bloggers link to *Wikipedia* for definitions of terms.

Myers (2010:32) also found that bloggers make extensive use of deixis; he gives the example of a Blogger referring to "*that movie*" which the reader can click on to find a link to what "*that movie*" is and find out more information about it. Bloggers also link to *Facebook* and *Twitter*. According to Myers, the blogger in these cases assumes that the reader follows *Twitter* and *Facebook* as well as the blog, thus a blog update may contain information regarding a post on *Twitter* or *Facebook* for example which contains a link to that information which the reader can go to.

Blogs also link to mainstream media for news information for example. Myers states that although Blogs were initially viewed as having the potential to undermine traditional news, in reality they actually rely on mainstream media and the online forms of traditional news (Myers 2010:33).

It is clear that link sharing is an important part of blogs. Blood (2000) argues that the early versions of blogs began through link sharing and that early definitions of blogs included dated entries, links and thoughts on personal websites. The ways in which bloggers use links include embedding the URL, phrase, name, title, number, word, deictic expression, brand name, image or quotation (Myers 2010:34). Myers identified six different functions of links in blogs. First, links can provide more information about what is discussed in the linked text. Second, they can provide evidence for the claims in the linked text. Third they can give credit to whoever gave them the information. Links can also lead to action, for example giving to charity. Myers also identified links as having the function of solving a puzzle and finally having the function of presenting different information, for example in irony. The blogging publishing tools allow the insertion of links easily thus bloggers have used the various methods of linking to present the information in the way they desire.

During the 2000s, weblogs have increased in volume and have become a form of mass self-expression on a variety of topics. However, due to the individual and idiosyncratic nature of blog writing, it is difficult to characterize the language of blogs as all conforming to one form or all belonging to one register. Therefore, before going further, it is essential to explain the terms 'register' and 'genre' in order to clarify the distinction between them and understand which term can be used in the discussion of weblogs.

Lee (2001) claims that the terms *genre* and *register* overlap and are often used interchangeably (Lee 2001:6). "One difference between the two is that genre tends to be associated more with organization of culture and social purposes around language… whereas register is associated with the organization of situation or immediate context" (Lee 2001:6-7). Lee also states that

> Register is used when we view text as… variety according to use [whereas] genre is used when we view the text as a member of a category: a grouping according to purposive goals, culturally defined…Genres are categories established by consensus within a culture and hence subject to change as generic conventions are contested/challenged and revised (Lee 2001:10).

Thus for the purposes of this paper, the term *genre* will be used in relation to weblogs. It is important to point out that there is some contention between scholars regarding weblogs being categorized as a genre. Nonetheless, Herring et al. (2004) have established that blogs are a distinct genre because they conform to Miller's definition of genre as "typified rhetorical action based in recurrent situations" (Miller 1984:159 cited in Herring et al. 2004:24). Herring et al. (2004:24) based their study on "the assumption that recurrent electronic communication practices can meaningfully be characterized as genres". Swales (1990) defines genre as "a class of communicative events" (Swales 1990:45) and argues, "the principal criterial feature that turns a collection of communicative events into a genre is some shared communicative purpose" (Swales 1990:46) which is "recognized by the expert members of the parent discourse community" (Swales 1990:58). This notion of a shared communicative purpose, which lies at the heart of Swales's definition of genre, was identified by Herring et al. in their study of blogs. Herring el al. found that all the blogs shared a common purpose regardless of type, which was expressing the author's opinion on matters of interest. Moreover, 90% of the blogs they examined were maintained by a single person, which led to the conclusion that "private individuals create blogs as a vehicle of self expression and self empowerment" (Herring et al. 2004:11). Herring et al. also suggest that weblogs are neither unique nor reproduced entirely from offline genres but constitute a hybrid genre that draws from multiple sources including other Internet genres. In addition, Herring et al. (2004:24) argue that that weblogs also have "similar structures, stylistic features, content and intended audience" and are culturally recognized and thus should be recognized as a genre.

In terms of differentiating between *genre* and *text type*, Biber (1988) provides the following description:

> Genre categories are determined on the basis of external criteria relating to the speakers purpose and topic, they are assigned on the basis of use rather than on the basis of form (Biber 1988: 170 cited in Lee 2001:38)

Thus according to Biber, text types are determined on the basis of form whereas genre is determined on the basis of purpose and topic. Lee (2001) states that there is some contention amongst researchers surrounding the use of the word *topic* in Biber's quote above. However, what is relevant for this discussion is that the "two texts may belong to the same text type (in Biber's sense) even though they come from two different genres because they have similarities in linguistic form" (Lee 2001:39).

In terms of categorizing blogs, there appears to be a consensus that the major distinction between blogs revolves around the personal the thematic or topic dimensions. This distinction is content based and provides a platform from which this study can investigate the linguistic features (the form) of the blogs.

Consider now figure 1. below which shows the first screen home page of a typical blog from the corpus.

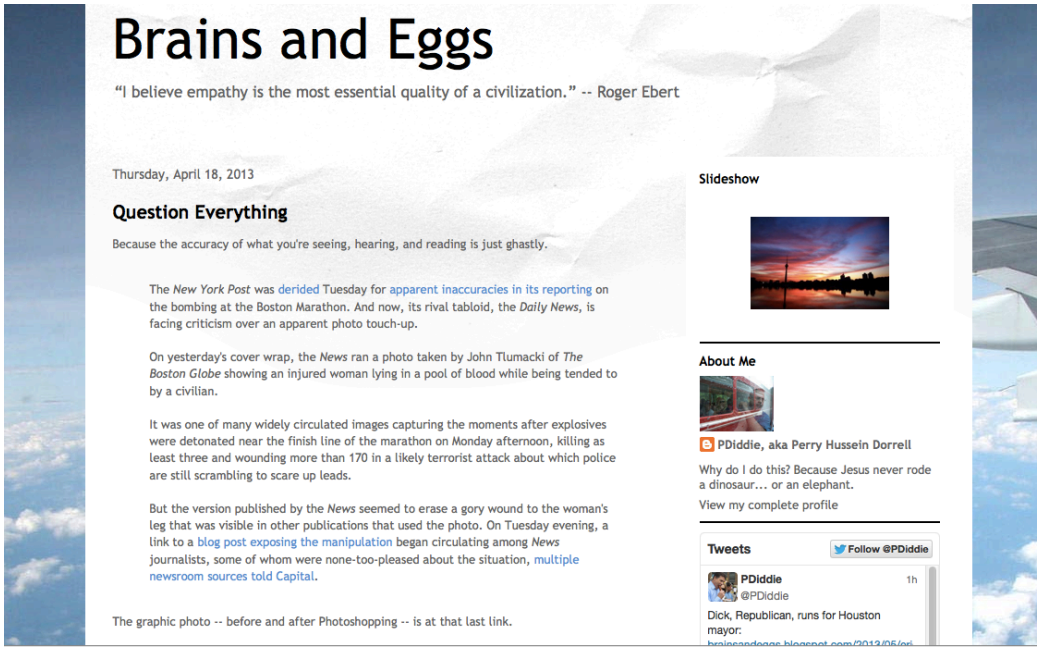

Figure 1. A sample blog

As is clear from the figure above, the title of the blog and the date of the blog post are typically presented at the top. On the right hand side in this example is the *about* section. This section varies from blog to blog with some positioning at the top, bottom, left or right of the page. Older entries are found at the bottom on the page and links to social media such as *Twitter* are also evident here on the right side of the page. This is optional and not all blogs have this feature although it is increasingly becoming more common.

Although Herring et al. (2004) found that the majority of blogs were created and maintained by a single individual, this study found that the majority of *Technorati's* higher rating blogs were multi-author blogs. The writers of the blogs examined provided a name and there was a means for contact as well in the form of email or the comments section. In Herring et al.'s (2004) study, filter blogs accounted only for 12.6% of their sample. However, with the *Technorati* directory categorizing blogs according to topic, it appears that in the nine years that have followed the publication of that work, filter blogs have increased in number and 7 out of the 8 categories examined in this study are filter (thematic blogs).

After understanding the basic structure and the main categorizations of blogs, the following section provides a discussion on formality and contextuality.

3.2. Formality and contextuality

3.2.1. Formal language

> By Formality we mean the use of technical, elevated or abstract vocabulary, complex sentence structures and the avoidance of the personal voice (I, you) (Coffin et al. 2003: 28)

The above extract is from the textbook *Teaching Academic Writing* and describes what is meant by formal language required in academic writing. The same textbook also provides the following characteristics for formal writing:

1. High lexical density
2. High nominal style
3. Impersonal constructions
4. Hedging and emphasizing

Texts which have a high frequency of verbs and personal pronouns are expected to be less formal than texts which use nominalizations. In addition, academic texts (especially the hard sciences) require objectivity and a degree of impersonality thus the use of the active voice is generally omitted in these texts.

Leech & Svartvik (2002:30) define formal language "as the type of language we use publicly for some serious purpose, for example in official reports, business letters, regulations and academic writing". Moreover they state that most of the vocabulary in formal language is of Latin and Greek origin whereas informal language is characterized by words of Anglo-Saxon origin. Leech and Svartvik (2002: 31) also state that "the difference between <formal> and <informal> usage is best seen as a scale". Informal language is defined as " the language

of ordinary conversation, of personal letters, and of private interaction in general" (Leech & Svartvik 2002: 30). The notion of active voice appears to be significant in informal language whereas in formal language an impersonal style is generally adopted. In terms of the formality scale it is generally regarded that casual conversation is at the lower end of the formality scale whilst academic writing and ceremonial speeches are at the high end of the formality scale.

Turning to CMC, because of its hybrid nature consisting of both spoken and written characteristics, in addition to the presence of different modes of CMC, its positioning along the formality scale varies according to the mode (Knight et al. 2013: 132). "Chat" for example is considered to be the most similar to conversation and thus informal whereas webpages are closer to written language and thus more formal. Tagg (2009) and Ling (2003) examined SMS messages and found them to be personal and contain a sense of immediacy. Tagg (2009:17 cited in Knight et al. 2013:132) argues, "The informal and intimate nature of texting encourages the use of speech-like language ". Yates (1996) and Crystal (2001) also recognized that CMC consists of characteristics of both spoken and written language. Baron's study of email found that although email is written, it is used "for typically spoken purposes" (Baron 1998:36 cited in Knight et al. 2013:132). Knight et al. also identified that the "levels of formality across e-language as a specific genre […] is something that remains under-explored in corpus -based analyses of real life data" (Knight et al. 2013: 132).

3.2.2. Contextuality and Formality

Heylighen and Dwaele (2002) claim that all communication refers to context to some degree and in some situations context will be more prominent than other situations (Heylighen and Dewaele 2002:2). They cite the anthropologist Edward T. Hall's distinction of two types of contexts: *High context* and *low context*. High context refers to situations where the context plays a vital role in understanding the communication, as the communication itself is implicit. Low context refers to situations where the context plays a minimal role in understanding the communication, thus the communication itself is explicit. This distinction was essentially made to differentiate between cultures although Heylighen and Dewaele have developed this further and incorporated it into the measure of formality (F-score, See section 3.2.3. below) by proposing that grammatical categories of words have different degrees of context dependency. Expressions that are context dependent or contextual are thus dependent to some degree on the context and are called deictic expressions or deixis. These types of expressions are "ambiguous when considered on their own, but where ambiguity can be resolved by

taking into account additional information from the context" (Heylighen and Dewaele 2002: 4), for example *I*, *me*, *he*, *she*, now, *then,* etc. Not only does contextuality incorporate the notion of deixis but it also includes implicature and anaphora. The following extract from Heylighen and Dewaele explains this further:

> The *context* of an expression can be defined as *everything available for awareness which is not part of the expression itself, but which is needed to correctly interpret the expression.* (Heylighen and Dewaele 2002:4)

Context is an important aspect of formality. In formal language, sharing of context is minimal whereas informal language maximizes on contextuality. For this reason, written language, where there is no direct interaction between interlocutors, contains less contextual information than casual conversation. Heylighen and Dewaele juxtapose contextuality with formality. Formal language, they argue, is explicit and avoids context dependency and ambiguity. Thus formal language does not rely on background knowledge or assumptions; rather these are explicitly stated in formal expressions. As a consequence, they claim that formal expressions are clearer and chances of misinterpretation of formal language are lower than informal language.

This study examines the language of blogs by first using the measure of formality called the F-score (see section 3.2.3 below) to determine contextuality and then examines the individual markers of formality identified above to study blog language formality and investigate whether blog topic affects blog language or not.

The level of language formality is an important aspect of language use. The same content can be expressed using very different writing styles ranging from the very formal to the very informal. However, what exactly constitutes formal language and how can it be determined?

Heylighen and Dewaele (1999), who developed a measure of formality, argue that although an intuitive determination of formality can be made, "a clear and general definition of formality is not obvious" (1999:1). Thus they set out to determine an empirical measure of formality known as the F-score. They claim that there are two types of formality: *surface formality* which can be summarized as the attention to form for the sake of the form itself, and *deep formality* which is "Attention to form for the sake of unequivocal understanding of the precise meaning of the expression". Heylighen and Dewaele (1999:2). Deep formality, they argue, is universal and more significant since surface formality will generally follow from deep formality.

Heylighen and Dewaele (1999) argue that "an expression is formal when it is context-independent and precise (i.e. non-fuzzy), that is, it represents a clear distinction which is invariant under changes of context" (Heylighen and Dewaele 1999: 8). They do however concede that formality is a relative concept and that all linguistic expressions are situated on a continuum between extreme formality and extreme informality, which is influenced by the personality of the producer of the linguistic expression and the situation in which the linguistic expression is produced.

Heylighen and Dewaele (1999:8) maintain Heylighen's (1993) observation that formal language can be "extend[ed] over wider contexts: more people, longer time spans and more diverse circumstances". In addition, formal expressions need more planning and attention in order to be produced. The lack of context in formal language means that formal expressions use a higher frequency of nouns necessary for making information explicit. Informal language on the other hand, can convey the same information with shorter and more common words. Contextual expressions are shorter and more direct, mainly because of the shared context. The shared context also means that informal language does not have the same need for precision as formal language. Moreover, contextual expressions are involved and interactive and non-verbal cues aid in making informal language understood whereas formal expressions are generally detached and impersonal (Heylighen and Dewaele 2002:5).

Deixis[2] plays an important part in the determination of formal and informal language because it varies according to context. This is because words with a deictic function refer to the context. Yule (1996) identifies three types of Deixis: person deixis, which refers to people (e.g. *I, me, he*), spatial deixis, which refers to place (e.g. *here, there*) and time deixis (e.g. *now, then*). However, Heylighen and Dewaele (1999) present a fourth type of deixis identified by Levelt (1989:45) as discourse deixis (e.g. *therefore, however*,). Discourse deixis includes anaphoric reference in addition to interjections such as "ooh", "well", "OK" "(Heylighen and Dewaele (1999:36). Discourse dexis is an indicator of both formal and informal texts.

Discourse deixis is also called text deixis and according to Levinson (1983) "concerns the use of expressions within some utterance to refer to some portion of the discourse that contains that utterance (including the utterance itself)"(Levinson 1983:85). Some of the expressions provided by Levinson (1983:87) to exemplify discourse deixis are: *but*, *therefore*, *in conclusions*, *to the contrary*, *still*, *however*, *anyway*, *well*, *besides*, *actually*, *all in all* etc.

[2] Deixis originally from Greek essentially means, "pointing via language" (Yule: 1996:9).

These discourse deictic expressions according to Levinson indicate that the utterance that contains them is a continuation of previous discourse (Levinson1983: 88).

Although Heylighen and Dewaele do not make a distinction between discourse deixis and anaphoric reference in their categorization, Levinson (1983) clarifies the difference between the two notions. Anaphora he states "concerns the use of (usually) a pronoun to refer to the same referent as some prior term"(Levinson 1983:85). Levinson further explains "deictic … expressions are often used to introduce a referent and anaphoric pronouns used to refer to the same entity thereafter"(Levinson 1983:86). Although this distinction is not significant for the F-score, it is important nonetheless to clarify that a distinction does exist.

Turning to non-deictic words, Heylighen and Dewaele state that they are not generally affected by change in context. Most nouns and adjectives are non-deictic words. On the other hand, pronouns, adverbs and interjections are context dependent words and inflected verbs are also considered deictic because they may refer to a certain time through their tense and a certain person or object through their inflection in addition to direction such as *come* and *bring* etc. As a consequence, formal language shows a higher frequency of nouns, which do not contain contextual information, whilst informal/contextual language will favor the use of verbs, which carry contextual information.

3.2.3. The F-score

The F-score proposed by Heylighen and Dewale in principle divides word classes into deictic and non-deictic categories. The word classes in the deictic categories are considered informal/contextual whereas the word classes in the non-deictic categories are considered formal. As stated above, nouns are typically non-deictic and thus considered formal whilst adjectives, because of their association with nouns, are also placed in the non-deictic category. Articles are also non-deictic because they co-vary with nouns. On the other hand, pronouns, which are clearly deictic, are considered markers of informality/contextuality as are verbs, which as stated above can contain deictic markers. At the same time, adverbs, due to their association with verbs, are also placed in the deictic category. In addition, interjections are also placed in the deictic category due to their frequency in more informal styles. Finally it must be noted that conjunctions have been omitted from the F-score as they are deemed to have no influence on formality. Heylighen and Dewale (1999) state:

> Conjunctions, which have no reference, neither to an implicit context, nor to an explicit, objective meaning, do not seem to be related to the deixis or formality of an expression, but only to its structure. Therefore, they are not put in either category. (Heylighen and Dewale 1999:13-14)

The F-score formula as proposed by Heylighen and Dewale (1999) is presented below:

F = (noun frequency + adjective freq. + preposition freq. + article freq. – pronoun freq. – verb freq. – adverb freq. – interjection freq. + 100)/2

Thus the F- score then is a measure based on the frequency of certain word classes in a text. This frequency is calculated by counting the occurrence of the word classes in the formula above in each blog using the concordance program *Wordsmith* and computing it as a percentage. The F- score ranges between 0-100; the higher the F-score, the more formal the language will be. Nouns, adjectives, articles and prepositions are more frequent in formal texts whilst pronouns, adverbs, verbs and interjections are found in more informal styles. Hence the conversation register investigated by Biber et al. (1999) found high frequencies of pronouns, verbs and interjections in their informal conversation register.

Heylighen and Dewaele (1999) state that although this measure of calculating formality is "coarse grained" (1999:1) it has been shown to distinguish between formal and informal genre in Dutch, French, Italian and English. Their results found that the frequency of formal categories (nouns articles adjectives and prepositions) increases with the increase in formality while the frequency of deictic categories (pronouns, verbs, adverbs, interjections) decreases. Heylighen and Dewaele (1999:16) state, however, that the overall result presented by the F-score is more reliable than a single word class category (e.g. examining only pronouns).

This measure of formality however should be taken as a measure of contextuality rather than a measure of formality in the traditional sense. That is, the F-score provides information with regards to the level of context dependency of a text. It provides a text's "contextuality versus formality" (Nowson 2006:50). Nowson states that although "the word formal is traditionally used in opposition to informal" (Nowson 2006:51), when discussing the F-score it should be clear that formality is used in opposition to contextuality. Nowson (2006: 51) also argues that "a lower F-score only implies greater contextuality". The F-score then should be taken as an indication of how context dependent a text is. Indeed this appears to be the intention of Hyelighen and Dewaele who in a subsequent (2002) publication on the F-score named their study '*Variation in contextuality of language: an empirical measure'* rather than the original (1999) title 'Formality of language: definitions measurements and behavioral determinants'. Nowson (2006) adopted the use of the F-score to investigate

individual differences within genres and for this reasons this study has also adopted this measure.

The above discussion has provided an overview of the most important concepts in this study and has explained the notions of formality and contextuality as well as the F-score. Moving on to section 3.3 below, an outline of the most prominent research on the language of blogs is presented.

3.3. Previous research

This section deals with previous research. Please note that the research has not been discussed in chronological order but in the order that suits this study best.

3.3.1.Herring and Paolillo (2006)

In terms of the effect of blog genre on language, Herring and Paolillo (2006) examined the effect of gender and genre on the language of Personal blogs and filter (thematic) blogs. Their research question was "whether gender or genre is a stronger predictor of linguistic variation in weblog writing" (Herring and Paolillo 2006: 444). In their study of gender and genre variation in weblogs they found that the genre of the blog had a more significant impact on the writing style than author gender. The features examined were characterized into two categories: female preferential and male preferential categories. The female preferential features are mainly personal pronouns, whereas the male preferential features are determiners, demonstratives, numbers and the possessive pronoun *its*. Their results showed that overwhelmingly blog genre was a stronger predictor of linguistic variation and that "diaries favored female preferential features whilst filter blogs favored male preferential features" (Herring and Paolillo 2006:447). They suggested that the blog genre appears to be gendered in terms of linguistic features and that diary blogs made frequent use of first person references whilst in contrast filter blogs made frequent use of third person references. They concluded by stating, "weblogs are not a uniform genre" (Herring and Paolillo 2006: 455) and ascertained that more research needs to be made in this area. The linguistic properties they examined did not allow for gender prediction, however Herring and Paolillo (2006) argue they can be seen as genre features that distinguish interactive language which is assumed to be female from informative language which is assumed to be male. Genre effects, they argue, could be mistaken for gender effects.

3.3.2. Nowson et al. (2005)

Nowson et al. (2005) investigated the language of personal blogs using the F-score proposed by Heylighen and Dewaele (1999). They first compared the blog corpus they collected with sub-corpora from the British National Corpus (BNC). Then they examined the effect of individual personality differences on the blogs' formality/contextuality. They found that gender and agreeableness had the greatest effect on contextuality.

The blog corpus was collected by asking bloggers to complete a socio-biographic questionnaire and each blogger was asked to submit blogs they had written a month prior to taking the questionnaire. Nowson et al. (2005) calculated the F-score of 17 BNC genres including both spoken and written material in order to place blogs on a scale. In addition, they also calculated the F-score of an email corpus previously collected. The results are presented in the following table taken from Nowson et al. (2005: 1668):

Table 1. Average F-score of selected genres from BNC

Genre	Ave F
Sermons	42.4
Lectures on Social Science	44.3
Unscripted Speeches	44.4
Fiction Prose	46.3
Personal Letters	49.7
Sports Mailing List E-Mails	50.0
Scripted Speeches	53.0
School Essay	53.2
Biography	56.3
Non Academic Social Science	56.9
Nat Broadsheet Social	57.5
Professional Letters	57.5
Nat Broadsheet Editorial	58.1
Nat Broadsheet Science	60.0
University Essays	60.3
Academic Social Science	60.6
Nat Broadsheet Reportage	62.2

It is clear from the table above that the spoken genres had a lower F-score than the written genres. Nowson et al. (2005) also situated their email and blog corpus in comparison to the

BNC sub genres. The results are presented in the table below from Nowson et al. (2005:1668).

Table 2. F-scores of the email and blog corpus in comparison to the BNC genres.

Genre	Ave F
Sports Mailing List E-Mails	50.0
E-Mail Corpus	*50.8*
Scripted speeches	53.0
School Essay	53.2
Blog Corpus	*53.3*
Biography	56.3

Nowson et al.(2005) conclude that the ordering of the genres in their study is similar to the ordering of the genres in Bibers (1988) MD analysis of spoken and written English where he ranked the genres based on the involved versus informational factor. The Blog corpus in Nowson et al. (2005) had a higher F-score than the email corpus, scripted speech and school essays. Nowson et al. state that the blog corpus had a higher F-score than the email corpus because the email corpus was selected by instructing participants to write emails to people they know whilst the blog corpus was collected from web published material where the blogger was unlikely to know the reader of the blog and hence have a less of a shared context in comparison to the emails which were written to friends.

3.3.3. Teddiman (2009)

One of the prominent pieces of research regarding blogs and contextuality is by Teddiman (2009) who analyzed online diaries (Personal blogs) using the F-score formula suggested by Heylighen and Dewaele (1999). A diary corpus and a corpus of diary comments were collected and analyzed for formality. In addition, the results were compared with previously collected F-scores on similar types of data by Nowson et al. (2005).

Nowson et al.'s (2005) study used the F-score to determine the formality score of a blog corpus compared to a subset genre of the British National Corpus (BNC). The blog genre in their study had an F-score of 53.3 making it more formal than email and school essays, but less formal than written biographies. Teddiman's study expands on that by adding a corpus of diary comments. In addition, Teddiman focuses on the word classes, which are considered to

bias contextuality such as pronouns. Moreover, Teddiman (2009:331) claims that since "the F-score does not distinguish between category members" an investigation into the individual word classes might be useful in further text categorization.

The F-scores for both the diary corpus and the comments corpus in Teddiman's study were 55.5, which was slightly higher than the F-score reported by Nowson et al. (2005) at 53.3, but does not affect their ranking and both the diary corpus and corpus of comments fell between school essays and biographies when compared to the F-scores calculated from data in the BNC. The blog corpus shared many features of more formal writing for example the relative number of nouns and prepositions used by the authors. However, Teddiman found that despite the diary corpus and the comments corpus having the same F-score, they do not share the same linguistic features, especially with regards to pronouns. Despite the overall pronoun frequency being very close, frequency of the various individual personal pronouns showed some variation. In both corpora, *I* was the most frequent pronoun 37 times per thousand words in each and showed patterns close to conversation corpus of the BNC. The significant difference between the two corpora was in the frequency of the second person pronoun *you*, which was 20 times more frequent per 1000 words in the diary comments than the diary corpus. The difference in the frequency of the possessive pronoun *your* was even more pronounced between the two corpora. The comments corpus displayed a frequency very similar to the spoken sub corpus of the BNC of 4.1 per 1000 words compared to 3.6 per 1000 words in the spoken sub corpus. On the other hand, the diary corpus showed a frequency of 1.6 per 1000 words, which indicates high involvement and suggests that even when the F-score is the same, an examination of the frequencies of the different linguistic features can show significant results.

Teddiman argues that although the blog comments corpus is more formal and less contextual than conversation it does show patterns that are similar to speech in terms of certain pronoun uses such as *you* and *your*. Teddiman concluded by stating that the F-score proposed by Heylighen and Dewaele can be used to accurately distinguish between genres; however, in order to understand why certain genres are different a closer look at the linguistic features is needed. Moreover, Teddiman maintains that first person personal pronouns are often considered to be markers of an interpersonal focus as identified by Biber (1988: 225), a factor more closely related to conversation, and by extension, to contextuality rather than formality and adds that the results "suggest an interesting interplay between categorical frequencies and relative genre similarities" (Teddiman 2009:333). This adds further weight to

Nowson's (2006) claim that the F-score should be taken as a measure for contextuality rather than formality in the traditional sense.

Taking into account Teddiman's (2009) findings, this study does not stop at the F-score but also examines the pronoun category in detail to see if there are differences in the usage of the different pronouns.

3.3.3. Daems et al. (2013)

Another recent influential study on the language of weblogs is by Daems et al. (2013) who carried out a multi dimensional analysis (MD) of weblogs based on the MD approach proposed by Biber (1988). Daems et al. ascertain that linguistic investigation into the language of blogs has been sparse and that weblogs, despite their immense worldwide popularity, have received little academic attention. They investigated whether the variable of blogger occupational background would have an affect on blog language. The two occupational backgrounds were the humanities and the exact sciences.

Using a blog corpus of 9 million words written by men in their twenties, Daems et al. (2013) carried out an MD analysis to identify the relations between linguistic features. MD analysis is used to examine a text for linguistics patterns where co-occurring patters in a text are grouped together into factors that represent particular functions and suggest functional variation. Daems et al. (2013) categorized their results into four dimensions as follows:

1. Factor1: Narration versus Instruction
2. Factor 2: Formal versus Casual
3. Factor 3: Diary versus background story
4. Factor 4: Reflections versus Report

The table below (taken from Daems et al. 2013:11) shows the linguistic features for each factor.

Table 3: linguistic features for each factor in Deams et al. (2013)

Factor 1	Positive	Third person pronouns, past tense, possessive personal pronouns, adverbs, particles, (coordinating conjunctions)
	Negative	Mean word length, proper nouns, (type-token ratio)
Factor 2	Positive	Subordinating prepositions and conjunctions, determiners, past participles, *wh*-determiners, adjectives, (gerunds and present participles, existential *there*)
	Negative	(Netspeak, interjections)
Factor 3	Positive	First person pronouns, personal pronouns, (third person singular neuter pronoun)
	Negative	Nouns
Factor 4	Positive	Second person pronouns, third person singular present tense verbs, modals, base form verbs, *wh*-adverbs, *wh*-pronouns

Each of the dimensions above consists of positive as well as negative linguistic features. Daems et al. state, "the presence of both positive and negative loadings on a factor should be regarded as two distinct sets occurring in a complementary distribution" (Daems et al. 2013:11). Thus the positive features are those that typically occur at one end of the scale in a particular factor whilst the negative features are those that are typically at the other end of the scale for that factor. Thus for example if we take factor 1, the positive features which include third person pronouns and past tense indicate narrative style. On the other hand the negative features of high mean word length, proper nouns and a high type token ratio are indicative of instruction and suggest informational style. The study concluded that on the one hand, the academic blogs contain features that confirm their academic status such as possibility modals and agentless passives whilst on the other hand they also consist of interactional features and personal qualities. Daems et al. (2013) conclude that Herring et al.'s (2004) description of blogs, as a 'hybrid genre' is plausible and indicate that blogs can have multiple uses. Herring et al. (2004: 11) claim that the "flexible hybrid nature of the blog format means it can express a wide range of genres, in accordance with the communicative needs of its users".

3.3.4. Grieve et al. (2010)

Grieve et al.'s factor analysis of a blog corpus of 2,261,520 words also provides a set of dimensions for blogs. The study, which used *globe of blogs.com* to select the weblogs, found the following factors:

1. Informational vs. 'personal focus'[3]
2. Addressee focus dimension
3. Thematic variation dimension
4. Narrative style dimension

The linguistic features for the factors are presented in the table 4. Taken from Grieve et al. (2010:308) Below. . Again like in the study by Daems et al. (2013) the features are presented as having a positive or negative loading on each factor.

Table 4. Linguistic features for each factor in Grieve et al. (2010)

Factor	Loading	Features
1	Positive	Prepositions, attributive adjectives, nominalizations, passives, WH relative clauses, *that* relative clauses, post nominal *to* clauses, post nominal *that* clauses
	Negative	Emphatics, first person pronouns, discourse particles, hedges, past tense, time adverbials, place adverbials, progressive verbs, *to* clauses with desire/intent/decision verbs, quantity nouns, activity verbs
2	Positive	Present tense, second person pronouns, *do* as PRO-verb, demonstrative pronouns, *be* as main verb, indefinite pronouns, WH questions, possibility modals, predictive modals, conditional subordination, necessity modals, mental verbs
	Negative	Prepositions, past tense
3	Positive	Demonstrative pronouns, emphatics, pronoun *it*, hedges, clausal coordination, adverbs, conjuncts, predicative adjectives, factive adverbs, likelihood adverbs
	Negative	Second person pronouns, nouns
4	Positive	*That* deletion, past tense, third person pronouns, adverbial subordination (other), *that* clauses with factive verbs, *to* clauses with speech act verbs, to clauses with modality/cause/effort verbs, communication verbs
	Negative	Nouns, attributive adjectives

For factor 1, the positive features such as nouns and passives suggest a nominal informational style and indicate high informational density. This type of style is typically characterized as formal. On the other hand, the negative features for factor 1 such as emphatics and first person pronouns indicate a verbal style with a high degree of involvement. The negative features in factor 1 roughly correspond to the deictic features with negative loadings on the F-score by Heylighen and Dewaele (1999). Texts that have a high frequency of the negative features in factor 1 are usually categorized as informal.

[3] *Personal* in factor 1 is in quotation marks because it is the only factor presented as a scale whereas the rest of the factors are not.

For factor 2 the positive features indicate interactivity and addressee focus thus blogs which had a high frequency of positive features in factor 2 directly referred to their audience whilst those a high frequency of negative features for factor 2 were less interactive.

Turning to factor 3, Blogs that had a high frequency of positive features imparted a conversational tone and consisted of a great variety of themes and topics where the blogger shifted quickly from one topic to the next. On the other hand, blogs that had a high frequency of negative features on factor 3 tended to focus only on one topic.

Finally, the blogs that scored high on factor 4 were narrative in style whilst the blogs which had a high frequency of the negative features had high informational density.

Grieve et al. (2010) conclude that through this MD analysis, two main blog types were identified: Personal blogs and thematic blogs and that "both of these blog types are characterized by a highly personal and conversational style which appears to be the standard blog voice" (Grieve et al. 2010: 320). The difference between them, however, is due to "the content of the blogs: blogs that focus on their authors lives are distinguished from blogs that focus on impersonal and informational topics" (Grieve et al.2010: 320). They also found a third type, which they labeled 'expert blog' which is written in a standard informational non-personal style, which is distinct from the typical blog style. This type of blog, however, is rare.

This overview shows that a number of features need to be taken into account when examining the language of blogs. The factor analysis is one of the ways in which blogs can be analyzed, however due to the extremely time-consuming nature of factor analysis and the need for computer programming skills, it is beyond the scope of this study. Nonetheless this overview has shown that factors 1 and 2 ('Informational vs. personal focus' and 'Addressee focus' dimension) correspond roughly to the linguistic features in the F-score.

3.3.5. Biber (1988)

This section provides an overview of Biber's (1988) seminal study of variation across speech and writing. This study is relevant due to its intricate analysis of the differences between speech and writing in addition to a factor and cluster analysis that was innovative at the time but remains equally pertinent today. This study is of significance to this current study mainly due to similarities that CMC has with both speech and writing.

The study is based on the premise that "there is no single parameter of linguistic variation that distinguishes among spoken and written genres" (Biber 1988:55). The MD analysis by Biber (1988) was based on 67 linguistic features identified from previous

linguistic research. Biber (1988) examined the functional dimensions of these linguistic features based on their co-occurrence distributions in the factor analysis. Through the factor analysis Biber (1988:115) identified six main factors. These are as follows:

1. Informational versus involved production
2. Narrative versus Non Narrative Concerns
3. Explicit versus Situation Dependent Reference
4. Overt Expression of Persuasion
5. Abstract Non Abstract Information
6. Online Informational Elaboration

Factor 1 proved to be the most significant. All the factors were divided into positive and negative features According to Biber (1988:105), the positive features in factor 1 were all "associated in one way or another with an involved, non-informational focus". The features that had the largest weight on factor 1 were private verbs, *that* deletion, present tense forms and contractions. These features suggest a verbal style and can also be considered interactive or involved.

Factor 2, 'Narrative versus Non Narrative Concerns' provides a distinction between texts that are primarily narrative in nature and texts that are non- narrative such as expository or descriptive texts (Biber 1988:115).

Factor 3 is the 'Explicit versus situation-Dependent Reference'. In this factor, the positive features are mainly WH-relative clauses in object position, pied piping constructions and WH-clauses in subject position (Biber 1988:102). According to Biber these three clause types share the same function of explicitness hence their co-occurrence in this factor. The negative features in factor 3 are time and place adverbials and adverbs. Texts that had a high frequency of negative features for factor 3 used situation dependent reference.

Factor 4,'Overt Expression of Persuasion' only has positive features. These include: infinitives, prediction modals and possibility modals (Biber 1988: 102). According to Biber, these features are markers of persuasion and texts with a high frequency of these features are in all likelihood persuasive texts either explicitly or via argumentation.

Factor 5 is the Abstract versus Non Abstract' dimension. This factor provides a distinction between texts that are abstract and technical in nature and texts that are less informational and abstract. The positive features in factor 5 include: conjuncts, agentless passives, past participle clauses, and BY- passives. Texts that have a high frequency of these

positive features are usually abstract in nature. Only type/token ratio had a negative loading on factor 5. This is however surprising because texts with a high type token ratio are typically characterized as highly informational texts. However Biber reasoned that this negative weight of type/token ratio on factor 5 indicates texts that are abstract and technical in nature reuse "a small set of precise technical vocabulary to refer to the exact concepts " (Biber 1988:112). However, it must be noted the method used by Biber (1988) for calculating type/token ratio is based on "counting the number of different lexical items that occur in the first 400 words of each text and then dividing by four" (Biber 1988:238). It is thus possible that the use of new measurement strategies such as standardized type token ratio might have given a different result.

'Online Informational Elaboration', factor 6 makes a distinction between elaborated informational discourse in high-pressure situations and consolidated informational and non-informational discourse. However Biber (1988:114) concedes that this factor needs further investigation.

To conclude this overview, Biber's (1988) study provides a number of factors that are useful for linguistic analysis of texts and a number of the features mentioned especially in factor 1 are incorporated in this study.

3.3.6. Knight et al. (2013)

Knight et al. (2013) examined the use of forms of hedging in e-language. The corpus used in their study was CANELC (the Cambridge and Nottingham e-language Corpus), a one million-word corpus of digital discourse taken from the British contributors to those posting to British websites in 2010 and 2011. It includes data taken from discussion boards, blogs, tweets, emails and SMS messages. Their main aim was to see how the different modes of e-language relate to Crystal's notion of the continuum of formality.

Knight et al (2013:133) state that spoken interaction is context specific and that a range of social, cultural and cognitive factors determine the meaning in spoken discourse. This is because in spoken language there is "shared knowledge about the immediate communicative context" (Knight et al.2013: l33) and for this reason contextual expressions such as *I*, *there* and *now* are frequently used. Written texts on the other hand are not as context bound, which means a "decrease in the use of contextual (deictic) expressions"(Knight et al. 2013:133).

One of the noticeable features of the data in the CANELC corpus is that despite it being asynchronous in nature, the researchers found that in most cases only a few seconds or

minutes passed between receiving a message and posting a reply which Knight et al. (2013) state "reduces the spatial and temporal distance between interlocutors" (2013:134). Thus they found a high frequency of deictic expressions such as the very high frequency of personal pronouns. This leads to a "shared digital space rather than physical space" (Knight et al.2013: 134).

Knight et al. (2013) examined the frequency of hedges in the CANELC corpus. The frequent use of hedges is often linked to formal rather than informal language. Knight et al. state, " given the tendency for writing to be more formal the level of hedging is generally higher for written discourse versus spoken discourse" (Knight et al.2013: 136). However, Knight et al. found that the rate of hedging in their data is inconsistent with typical rates in spoken and written discourse and argue that this result provides an argument for classifying e-language as its own distinct genre. Knight et al. found that "more immediate forms of e-language (e.g. SMS) are positioned closer to the spoken end [of the formality continuum] while emails and blogs are better positioned towards the more formal written end" (Knight et al. 2013: 149).

Knight et al. (2013) also found that the most frequent hedge in the CANELC is the adverb *just* followed by *really* and *only*. They also found that the "use of hedging in e-language shows some clear similarities with those used in more informal spoken discourse" (Knight et al. 2013:147). Their results indicate that the hedges *actually, just, you know, probably, quite, really, thing* are significantly underused in CANELC compared to the written BNC but significantly over-used compared to the spoken BNC sub corpus. The hedge *just* was significantly underused in the blog data and over-used in SMS messages. The hedge *you know* was also underused in the blog data as well as the message board data but over-used in the email and SMS data. In general, the hedges that were over-used in the CANELC corpus are *apparently*, *guess* and *maybe* although *guess* was underused in blogs but significantly over-used in SMS. The hedges *maybe* and *likely* were both underused in the blog data but over-used in the SMS data while *likely* in particular was also over-used in the *Twitter* data.

Knight et al. also divided the data into groups along a continuum of public to private. The categories are presented in the figure below taken from (Knight et al.: 138).

Topic / Genre Codes:		Topic / Genre Codes:	
	News, Media and Current Affairs	D	Music
A	Politics		Sports
	Business and Finance		
	Weather and the Environment		Celebrity news and gossip
		E	TV
	Culture, Literature and the Arts		Humour
B	Fashion		
	Teaching, Academia and Education		Health and Beauty
		F	Parenting and Family Life
	Technology, Computers and gaming		Personal and Daily Life
C	Hobbies and Pastimes		
	Travel		
	Cookery		

Figure 2. Categories in Knight et al. (2013)

Category A contains topics that are the most public and thus by association more formal whilst group F is the most personal and involved and thus by association the least formal. Categories B-E are in-between. They found the more public and the more personal topics had the highest frequency of hedges whilst the other topics were in between along the continuum. Knight et al. maintain that although spoken contexts often have a higher frequency of hedges than written contexts, both formal speech and writing employ a more frequent use of hedges in comparison to informal contexts. From their study, the category related to personal issues and thus by association similar to speech makes frequent use of hedging. In addition, Category A, which included public topics such as news and politics and thus associated with more formal topics, also displayed a higher frequency of hedges. However Knight et al. found "that there is no clear-cut relationship between the use of hedging in e-language compared to written and spoken genres of discourse" (Knight et al.2013: 149) and this they state adds further weight to the argument that CMC is a "distinct variety on the continuum of formality: between spoken and written discourse" (Knight et a. 2013:149).

4. Material

The present study is based on an analysis of 160 000 words from eight different blog topics: seven thematic (filter) blog categories and one personal blog category. The seven thematic

blog categories were selected from *Technorati* the word's largest blog search engine. These blogs were selected based on the following criteria: they had to be single-author blogs from bloggers who were not journalists. Thus blogs as part of newspapers and magazines were excluded. Moreover the blogs had to adhere to the topic. Although the *Technorati* blog directory provided assistance in choosing the blogs according to topic, the content of the blogs was checked before it was included. One of the difficulties in gathering the data was that the blogs listed higher up in *Technorati's* blog classifications were mostly multi author blogs or part of newspapers, magazines or websites and thus did not fully adhere to the traditional form of blog. In addition, a large number of blogs had editors. This problem was especially noticed in the topics of sport and entrainment. Once the blogs were selected for inclusion in the corpus, only the blog text was included and boilerplates were excluded. All the blogs were current blogs, and blogs which contained fewer than 10 entries were excluded in order to ensure currency of material and enable comparability. In addition, only blogs that were updated within the last month of collecting the data were included. Moreover, blogs that revolved mainly around photos and contained only a small amount of text were excluded. Furthermore, extended quotes and quoted passages for example from newspapers, were excluded from the blogs so that the extended quotes from other sources do not skew the results, as they are not representative of a typical blog text. Lists were also excluded such as the ingredient lists in the Food blogs and the items lists in the Art blogs.

For each topic category, 20 000 words were collected from 4 different blogs. This was in order to get the most representative sample possible. Thus 5000 words were collected from each blog. In total 28 blogs were used in this study. The blog topics are as follows: *Political blogs* which focus on politics, *Food blogs* which focus on food, *Family blogs* which focus on parenting and family related information*, Finance blogs* which focused and finances and money-related issues, *Sports blogs* which focused on sports and for this category 4 different sports were selected: hockey, tennis, basketball and baseball, *Art blogs* which focused on handcraft and sewing and finally *Celebrity blogs* which focused on celebrity gossip.

The Personal blogs were collected via a search on the blog directory bloggeries.com for *Personal blogs* because the *Technorati* blog directory did not have a personal blog category. To ensure comparability, the same method that was used in collecting the thematic blogs was used to collect the Personal blogs and 20 000 words from 4 different Personal blogs were collected.

5. Method

The blog corpus was constructed through a random selection of publicly available blogs hosted by Technorati.com. Technorati blog directory was searched for single author blogs and the data was collected between March 2013 and April 2013. The corpus represents 32 users, each text sample was 5000 words with the entire corpus containing 160 000 words. Boilerplates were not included in the word count. The data was then manually copied and pasted onto a text document ready for tagging using the Stanford part of speech tagger (Toutanova et al. 2003). The Stanford tagger uses the Penn Tree Bank part of speech tags. The tagged blogs were then saved as text only documents and were then searched for part of speech using *Wordsmith tools 5.0*. (Scott 2005) and the results were manually verified. The Stanford part of speech tagger, which was used by Daems et al. (2013), was deemed appropriate because it takes into consideration three words left and right of the tagged word to improve the tagging accuracy. However, it must be acknowledged that since the tagger is trained on "(semi-) controlled language production" (Daems at al. 2013:9), it may produce some errors when tagging blogs. Daems et al. (2013) estimated that the accuracy of the tagger for their blog corpus was at least 89% and that the lowest limit was in the case of unknown words. This percentage of accuracy in addition to manual examination of specific linguistic features is expected to enhance the reliability of the results.

The Stanford part of speech tagger was thus used to tag for parts of speech categories in order to incorporate them into the F-score formula. However, because of the difficulty in tagging a blog corpus due to irregularities in spelling and the presence of errors in blogs, the results were manually proofread and revised. In addition, in order to reduce the margin of error and calculation mistakes, Microsoft Excel was used for all calculations (such as the F-score formula and calculating totals) thus insuring accuracy. To calculate the F-scores of each blog, first the blog corpus was tagged, and then searched for each word class tag (e.g. nouns, verbs etc.) using *Wordsmith*. The frequency of that word class was then calculated as a percentage in Excel. When all the word class frequencies were collected, Excel was used to calculate the F-score formula.

F = (noun frequency + adjective freq. + preposition freq. + article freq. – pronoun freq. – verb freq. – adverb freq. – interjection freq. + 100)/2

Furthermore, the F-score was also calculated for the individual blogs within each topic. Thus

for each blog topic four different F-scores were calculated. This was to examine the degree of homogeneity within each blog topic and to ascertain whether the different blogs on the same topic had similar F-scores or if one blog had a higher or lower F-score that could skew the overall F-score result for that topic. The results for the individual blogs for each topic are presented after the initial analysis of the overall F-score for each topic.

In addition to the F-score, the other features that were examined in this study are: Type token ratio, key words, pronouns, contractions, hedges, downtoners, amplifiers, emphatics, discourse particles, private verbs, modal verbs, expletives and word and sentence length (see section 6). The Brown Family of Corpora was used in order to calculate key words and word and sentence length. It was chosen because it contains a wide variety of written texts. The second part of the method involves using the concordance program *Wordsmith* to search for the frequencies of the various linguistic features, which were then normalized to frequency per 1000 words. The results were then presented into tables and figures ready for analysis. The results were then used to compare which blogs had the highest frequency of each feature and which had the lowest frequency. Ellegård's (1978) study of the Brown corpus was used to compare the frequencies of the word classes in this study. By using these various tools, the margin of error is reduced and the results of this study can be considered reliable. Each of these features is integrated into the results section where they are used as a platform for presenting the results. These features were selected because of their prevalence in Bibers (1988) MD analysis. However, present tense verbs were not included due to the scope of this study. The results are presented in section 6.

6. Results

6.1. F-score results

Table 5: F-score for each blog according to topic

Topic	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
F-score	65.75	60.43	59.38	57.47	63.24	66.98	63.66	55.09

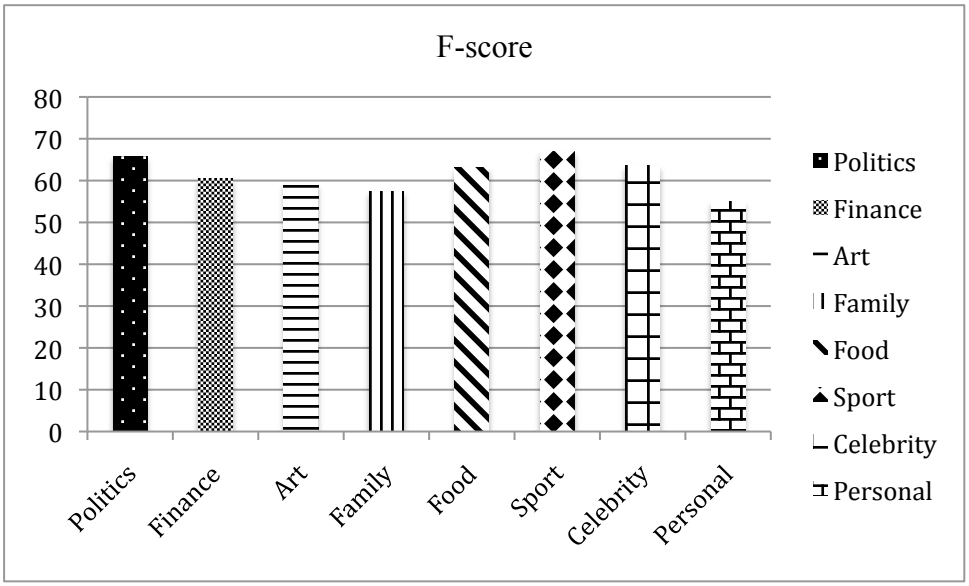

Figure 3. F –score for each blog.

Personal blogs scored the overall lowest F-score whilst the sport blogs scored the highest F-score closely followed by the politics and Food blogs. The F-score suggests that there is a difference between the different blog topics. The fact that the Personal blogs scored the lowest suggests that they are highly contextual whilst the fact that the Sports blogs scored the highest suggests they are more formal/informational. The F-score of the Personal blogs in this study (F-55.09) is slightly lower then the F-score for the Personal blogs in Nowson et al. (2005) F-53.3 and Teddiman (2009) F-55.5.

When comparing the F-scores in this study to the F-scores of the BNC subgenres investigated by Nowson et al. (2005), it becomes apparent the Personal blogs in this study fall between School Essays and Biographies; thus their positioning is the same as the blogs examined by Nowson et al. (2005) and Teddiman (2009) (see section 3.3.2). The Family blogs scored between Non-academic Social Science and National broadsheet social. The Art blogs scored between the National Broadsheet Editorial and National Broadsheet Science. The Finance blogs scored between University Essays and Academic Social Science whilst the Food, Celebrity, Political and Sports blogs all scored higher than the BNC genres presented in Nowson et al. (2005) which had a maximum F-score of 62.2. However turning to a more recent study, the blogs investigated by Lahiri et al. (2011) had an overall F-score of 65.24. These blogs were collected from the top 100 *Technorati* blogs (see appendix A for table of Lahari et al.'s blog F-score in comparison to the F-score of Forums, News, and Academic Papers). Lahirir et al.'s result is closer, in terms of F-score value, to the F-score of the political and Sports blogs in this study. It can be thus concluded that the F-scores of the

thematic blogs excluding the Family and Art blogs are rather high perhaps indicating that the writers of those blogs take a more formal less contextual approach to writing their blogs and assume less of a shared context with the reader. In order to understand these results further, a closer examination of the different word classes in each blog is presented.

6.2. Word classes for each blog

Consider now table 6, which shows the frequency of the different word classes per 1000 words for each blog topic.

Table 6. Frequency of the different word classes per 1000 words

Word class	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
Nouns	316.5	265.4	268.1	244.1	279.8	303.45	319.15	233.95
Adjectives	76.25	79.85	76.35	70.75	89.4	72.35	71.6	68.15
Prepositions	136.1	128.45	107.5	125.3	132.75	141.95	131.7	125.05
Articles	87.85	71	76.25	74.15	73.6	96.8	72.35	70.45
Pronouns	71.2	94.1	105.95	100.55	80.4	56.25	86.35	125.7
Verbs	177.75	176.95	186.5	195.55	163.9	162.35	178.45	200.65
Adverbs	51.3	64.75	56.25	58.45	66.05	55.75	55.9	67.05
Interjections	1.45	0.3	2.95	0.45	0.45	0.55	0.85	2.35

The table above shows that the Celebrity blogs followed by the Politics blogs and then the Sports blogs had the highest frequency of nouns. This is probably due to the reference to names. The Celebrity blogs mention different celebrities very frequently; the Political blogs reference the names of politicians whilst the Sports blogs make frequent mention of sports players' names. This no doubt has an effect on the frequency of nouns in the corpus. Across the board however nouns were the most frequent word class in all the blogs.

The verbs were most frequent in the Personal blogs indicating the active nature of the Personal blogs referring to activities and processes undertaken by the blogger. The family and the Art blogs also scored high on the frequency of verbs. The Sports blogs on the other hand had a low frequency of verbs especially in comparison to the Personal blogs 162.35 versus 200.65 per 1000 words. In fact, the Sports blogs had the lowest frequency of verbs out of all the blog topics.

The category of pronouns mirrors the results of the category of verbs. The Personal blogs had the highest frequency of pronouns overall followed by the Family blogs and the Art

blogs. This suggests that like the Personal blogs, family and Art blogs make frequent use of personal and deictic reference, which suggests that they are highly contextual. In contrast, the Sports blogs had the lowest frequency of pronouns and this difference is especially noticeable when compared to the frequency of pronouns in the Personal blogs 56.25 versus 125.7 per 1000 words respectively. Again this suggests that the Sports blogs are the least contextual whilst the Personal blogs are the most contextual.

Consider the figure below, which shows the frequency of the different word classes in each blog topic.

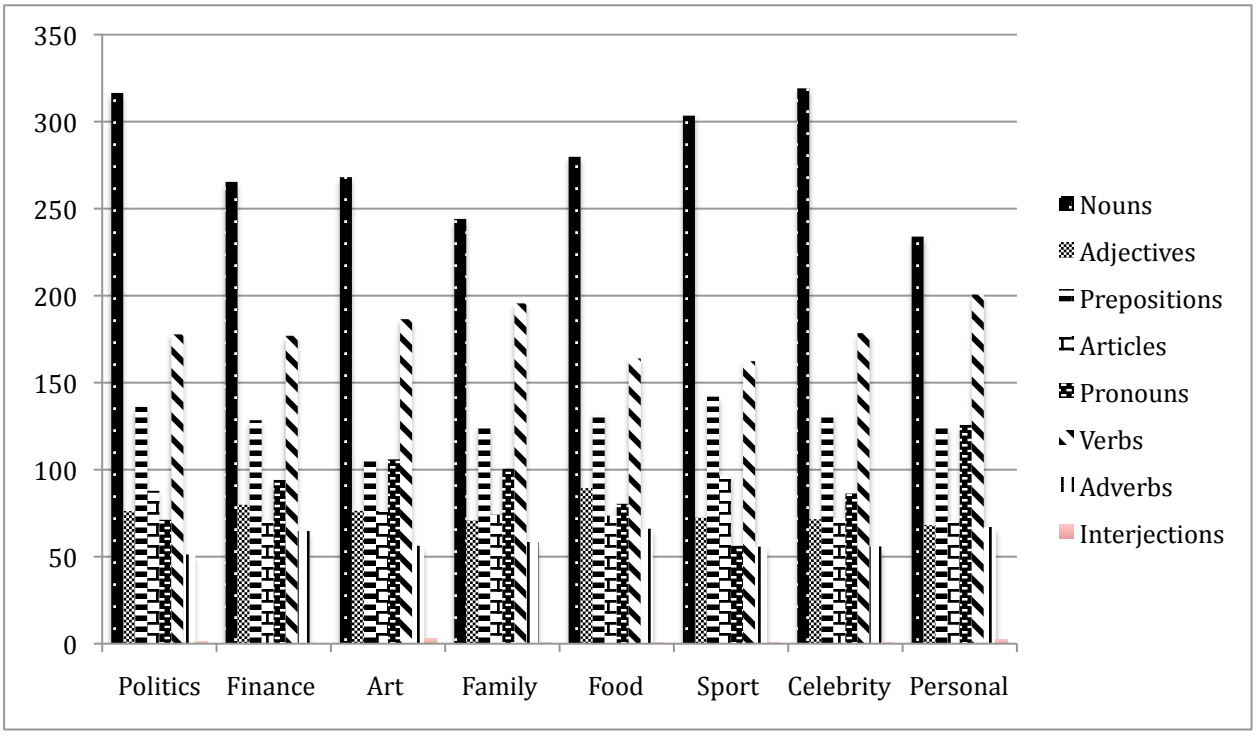

Figure 4 . Frequency of word classes per 1000 words for each blog topic.

From the figure above it clear that the frequency of nouns was more frequent in all the blog topics. Thus the blogs in general show a pattern that differs somewhat from conversation, which according to Biber et al.'s (1999) study had a slightly higher frequency of pronouns than nouns.

Ellegård's (1978) study of 128,000 words from the Standard Corpus of Present Day American English from Brown University is used as a frame for the results of this study. Four types of texts were examined: Popular Fiction labeled (N), Journalism labeled (A), Literary Essays labeled (G) and Science labeled (J). The material was divided into two parts in order

to assess the degree of homogeneity in each text type. The full table is presented in Appendix B. However, in order to be able to compare each category with the results from the present study more easily, an average was calculated for the two parts and used as a means of comparison. The table below shows the average frequency of the different word classes for each type of text.

Table 7. Average word class frequency per 1000 words from Ellegård (1978)

	Nouns	Adjectives	Prepositions	Articles	Pronouns	Verbs	Adverbs
Popular Fiction	216	48	100	90	168	228	74
Journalism	330	60	118.5	103	68	171	36
Literary Essays	253	87	126.5	103.5	114.5	177.5	55
Science	268.5	100.5	147.5	112	57.2	161.5	46

In the Brown corpus, 27% of all words were nouns. Thus nouns were the most frequent class of words. This result is similar to the results found in this study where for every blog topic; nouns were the most frequent word class. However, in the Brown corpus, Popular Fiction texts had a slightly lower frequency of nouns compared to verbs.

Verbs were the second most frequent class in the Brown corpus totaling 19% of all words. The Popular Fiction texts had the highest frequency of verbs with an average of 228 per 1000 words. This frequency is similar to the Personal blogs in this study, which also had the highest frequency of verbs with a frequency of 200.65 per 1000 words. The Science texts had the overall lowest frequency of verbs although Ellegård (1978) found that the frequency of the verb *be* in the science texts was higher than average.

Adjectives made up 7% of all the words in the Brown corpus. The scientific texts and the literary essays had the highest frequency of adjectives. In the present study, the Food and Finance blogs had the greatest frequency of adjectives. The Food blogs displayed a frequency similar to the literary essays with a frequency of 89.4 per 1000 words.

The adverbs were relatively small in the Brown corpus totaling only 5% of total words. The Popular texts had the highest frequency of adverbs whilst the journalistic texts had the lowest frequency. This result is mirrored by the present study where the adverbs were the least frequent word class excluding interjections in all the blog topics. The Personal blogs in

this present study had the highest frequency of adverbs suggesting similarities to the Popular Fiction texts whilst the Political blogs had the lowest frequency.

Personal pronouns in general made up about 10% of all the words in the Brown corpus. Personal pronouns including possessives were more frequent in the Popular Fiction texts and the least frequent in the scientific texts. In this study the Personal blogs had the highest frequency of Personal pronouns whilst the Sports blogs had the lowest frequency of personal pronouns.

Prepositions were more frequent in the scientific texts and least frequent in the Popular Fiction texts. In this study the Art blogs followed by the Personal blogs and the Family blogs had the lowest frequency of prepositions, which again suggest a similarity with the Popular Fiction texts, whilst the Sports blogs had the highest frequency of prepositions.

6.3. Type-token ratio

Type-token ratio can be used as a rough indicator of lexical density. The type-token ratio used in this study is the standardized type token ratio, which gives the type-token ratio per 1000 words. This gives a better indication of lexical density especially in larger corpora. The concordance program *Wordsmith* is used to calculate the standardized type-token ratio for each of the blog topics. Type-token ratio is considered as a rough indicator of lexical density, which is "a statistical measure of the relative frequency of lexical words and grammatical words in a stretch of text" (Flowerdew 2013: 29). Texts with a high type/token ratio are considered to be more informational than texts with low type-token ratio.

Table 8. Standardized type-token ratio in the blogs and the Brown Family corpus

	Politics	**Finance**	**Art**	**Family**	**Food**	**Sport**	**Celebrity**	**Personal**	**Brown family corpus**
Type token-ratio	53.73	42.95	44.38	45.44	47.72	46.17	51.87	47.22	40.62

The type token ratio is considered to be a rough indication of lexical density. From the table above it is evident that the Finance blogs had the lowest type token ratio, followed by the Art blogs and then the Family blogs. Although this result does not follow expectations, a closer inspection of the texts shows that the topics in these blogs are recurrent and thus a repetition of words is expected. The Sports blogs, which had the highest F- score, also had a low type token ratio. A closer look however shows that sports players' names as well as teams' names

are frequently repeated hence explain the low type token ratio. This is especially the case because each blogger focuses mainly on the team that they support thus the team and the players recur throughout. A low type-token ratio may also indicate that the same vocabulary is re- used to refer to the same concepts. Thus, in technical abstract texts, Biber (1988) found the type-token ratio to be low. The celebrity and the Politics blogs on the other hand had the highest type-token ratio, which may suggest a high assortment of topics in these blogs. The Politics blogs discussed a wide variety of political issues hence the use of a wider lexicon. The Celebrity blogs also discussed a large number of celebrities' lives, which is perhaps why they had a high type token ratio with the various celebrities' names. The low type-token ratio of the Brown Family of corpora perhaps indicates repetition in the corpora.

6.4. Key words

Key words are words in a corpus that have an unusually high frequency or an unusually low frequency in comparison to a reference corpus. The unusually high frequency key words are called positive key words whilst the unusually low frequency key words are called negative key words. The reference corpus used in this study is the Brown Family of corpora. The concordance program *Wordsmith* was used to calculate key words in the blog corpus. The key words were calculated by comparing the frequencies of all the words in the reference corpus with the frequency of all the words in the blog corpus. Frequencies that are higher or lower than usual are presented in the key word list.

The Brown Family of corpora was selected as a reference due to its breadth and the presence of a whole range of formality types in the corpus. The Brown Family corpus is composed of four corpora, two American English corpora: *The Brown University corpus of written American English (Brown)* and *The Freiburg- Brown corpus of American English (Frown)* and two British English corpora: *The Lancaster –Oslo/Bergen corpus (LOB)* and *The Freiburg-Lob corpus of British English (FLOB)*. Each corpus in the family of corpora consists of approximately one million words.

Table 9. Positive key words

Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
Obama	I	I	I	Minutes	Rangers	Her	I
Perry	Your	My	My	I	Seahawks	She	My
Benghazi	You	Blog	Kids	Bowl	pospisil	Mariah	Am
Whitehouse	Credit	Inspired	Kirsty	Butter	Season	Kardashian	Me

The table above shows the positive key words in each blog topic. The Politics blogs consisted of a higher than normal frequency of politicians' names. The American president's name Obama, for example, was the most significant followed by Perry referring to Rick Perry. Benghazi was also in the top three key words. This is because Benghazi was in the news during the time the blogs were collected thus it is expected that it should feature frequently in the Political blogs.

The Finance blogs consisted of a higher frequency than expected of the personal pronouns *I* and *your*. However, upon closer inspection of the Finance blogs it becomes apparent that the bloggers are providing their own personal insight on financial issues in their own lives. Thus for example one blogger describes living on one dollar a week. This sort of experience requires the frequent use of personal reference. The Finance blogs also saw the word *credit* as a positive key word, which is to be expected since the topic in question is financial.

The Art blogs had a high frequency of the personal pronouns *I* and *my*. This is because the Art blogs consisted on various explanations of artistic projects that the bloggers themselves have undertaken. Thus the bloggers make frequent use of the active voice and personal pronouns to show what they have made using art supplies or what they have embroidered or sewn, and in these contexts the use of personal pronouns is expected to be frequent. The word *blog* is also a key word due to frequent reference to the art blog made by the bloggers for example in directing the readers to return to the blog at a later time or to view another blog on a related topic. The fourth key word is the word *inspired* which considering the topic is art and inspiration is essential in artwork, it is not surprising that such a word is more frequent than expected.

In the Food blogs the word *minutes* is the most significant positive key word and this is because Food blogs consist of instructions for cooking and baking which require cooking and baking times thus the word *minutes* is expected to be amongst the most frequently used words in these types of blogs. The second positive key word is the personal pronoun *I*.

Perhaps the reason for this is that the Food blogs are written by individuals who cook and post their cooking online in addition to providing personal opinions on restaurants and certain foods. As a result, a higher than usual use of the personal pronoun *I* is to be expected in this context.

In the Sports blogs the most significant positive key words are team names. This is because the Sports blogs in this study mainly provide information about teams and players and describe the various sports being played without focusing too much on the bloggers own personal perspective. For this reason no personal pronouns featured in the top four key words.

The Celebrity blogs showed some interesting results. The top two positive key words were *her* and *she* which suggest that the Celebrity blogs present frequent information about females. A closer look at the Celebrity blogs shows that the bloggers are all female which may explain the preference for reporting on females more than males hence the more than expected use of *she* and *her*.

Finally, the Personal blogs provide the most striking result. All the positive key words are personal pronouns. This again supports the notion that Personal blogs focus on the blogger's own life and their internal thoughts and views hence the high frequency od personal pronouns especially in the first person.

Table 10. Negative key words

Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
Was	By	Which	The	Were	Its	Its	Its
Had	Were	Their	He	The	Which	Said	Which
She	She	In	Of	She	Of	Which	The
Her	He	Was	-	By	Said	Or	Of

Turning to the key words that were used infrequently in comparison with the Brown Family of corpora, it becomes apparent that in the Politics blogs, the verbs *was* and *had* were used infrequently, which perhaps indicates that the Political blogs focus on current events. In addition, the third person pronouns *she* and *her* were significantly underused in the Political blogs. This pattern of underuse of feminine pronouns was also noticed by Biber et al. (1999) who attributed the low frequency of feminine pronouns to traditional gender roles. The Food blogs also registered a low frequency of the feminine third person pronoun *she*. The Family blogs on the other hand registered obvious infrequent use of the masculine third person pronoun *he*. It could be that Family blogs focus on female related issues thus reducing the

need for frequent use of the masculine third person pronoun *he*. The Finance blogs had a significantly low frequency of both the feminine and the masculine third person pronouns. Third person pronouns are normally frequent in fiction, reported speech and descriptive narrative. This suggests that the blogs with a low frequency of third person pronouns are written in a more direct style and do not discuss other people frequently and thus have a lower need for third person pronouns.

6.5. Pronouns

Pronouns, and personal pronouns in particular are the biggest indicators of contextuality and involved style. The frequencies of the different types of pronouns are calculated for each blog topic and compared. The following types of pronouns are investigated: personal pronouns, demonstrative pronouns and possessive pronouns.

6.5.1. Personal pronouns

Personal pronouns are "function words which make it possible to refer succinctly to speaker/ writer, the addressee, and the identifiable things or persons other than the speaker/writer and the addressee" (Biber 1999:328). Personal pronouns were the most frequent type of pronoun in Biber et al.'s (1999) forty million word spoken and written corpus although their distribution varied across registers. Personal pronouns were significantly frequent in the spoken corpus and less so in the academic texts corpus. In this study, the frequency of all the personal pronouns was calculated for all the blog topics using the concordance program *Wordsmith* and computed to a frequency per 1000 words.

Table 11. Total frequency per 1000 words of all personal pronouns in each blog

Topic	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
Frequency	48.35	69.1	81.8	75.35	64	35.5	54.85	94

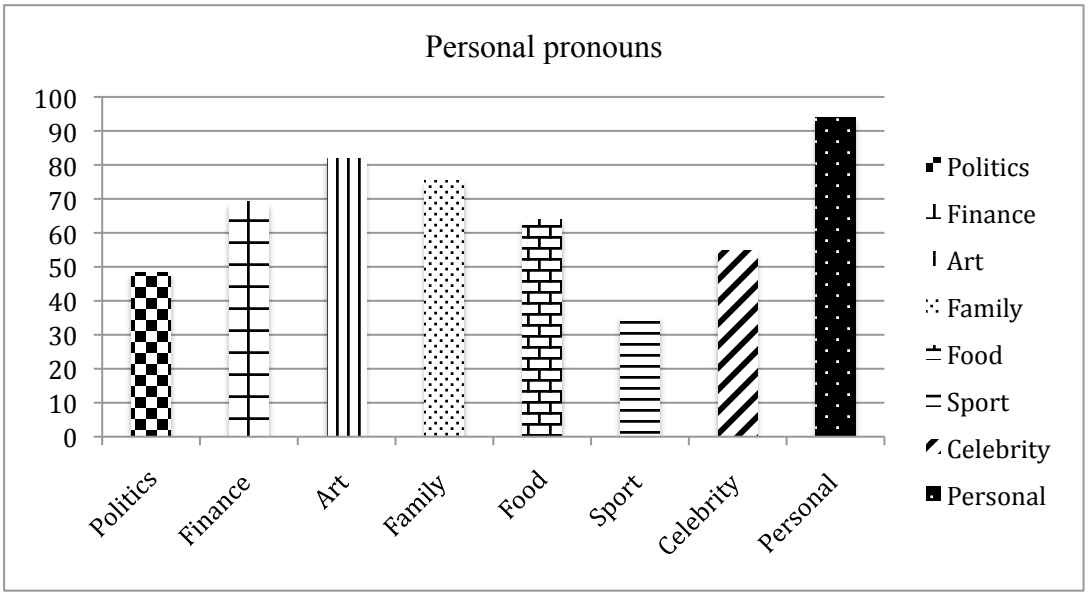

Figure 5. Frequency of personal pronouns in the different blog topics per 1000 words

The results above show that the Personal blogs by far had the highest frequency of personal pronouns. This indicates that the Personal blogs differ from the other thematic blogs based on the frequency of personal pronouns. The Art blogs followed by the Family blogs were the other two blogs with a high frequency of personal pronouns. The reason for this is that the Art blogs discuss art projects that the blogger has undertaken and thus require frequent reference to one's self. The Family blogs also make frequent reference to one's own self and other members of the family in addition to parenting experience. This requires the use of personal pronouns hence the high frequency in the results. Perhaps the surprising result is the Sports blogs, which had a very low frequency of personal pronouns. If we compare the frequency of the personal pronouns in the Personal blogs 94 per 1000 words to the Sports blogs 35.5 per 1000 words we can see that the Sports blogs used personal pronouns 61.5 times less per 1000 words than the Personal blogs. Upon closer inspection it becomes clear that the Sports blogs focus on describing the sports game or team and the actions of the players and thus make frequent use of nouns to refer to the players names or teams which reduces the need for the use of personal pronouns. Consider now table 12. which shows the frequency of the different personal pronouns in each blog topic per 1000 words.

Table 12: The frequency of the different personal pronouns in each blog topic

Personal pronouns	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
I	7.15	23.9	34.35	20.55	23.75	8.1	9.85	40.3
Me	1.35	2.6	2.75	3.85	2.8	1	1.35	7.6
You	6.3	17.45	12.45	8.5	11.2	2.1	5.4	10.45
It	10.55	13.1	19.7	12.3	15.9	5.3	9.75	13.05
He	9.7	1.45	0.3	3.35	0.1	10.65	4.35	2.2
She	0.6	1.1	1.3	5.15	1.3	0	13.55	5.8
They	4.6	3.05	2	6.35	2.3	0	2.95	3.5
Them	1.6	1	3.05	3.15	2.3	10.65	0.75	1.6
We	2.9	3.8	4.2	7.6	2.85	0	1.25	6.3
Us	0.75	0.9	0.7	1.45	0.65	0	0.5	0.65
Himself	0.55	0	0	0.25	0	0	0.15	0
Herself	0	0	0	0.05	0	0	1.25	0.1
Itself	0.25	0.05	0.15	0	0	0	0.1	0.05
Themselves	0.15	0.05	0	0	0.05	0	0.05	0.05
Yourself	0	0.2	0.2	0.3	0.15	0.05	0.05	0.1
Total	46.5	68.65	81.15	72.85	63.35	37.85	51.3	91.75

Consider now figure 6. which clearly shows the frequencies of the different personal pronouns in each blog topic. From this figure the results from the table above are more easily interpreted.

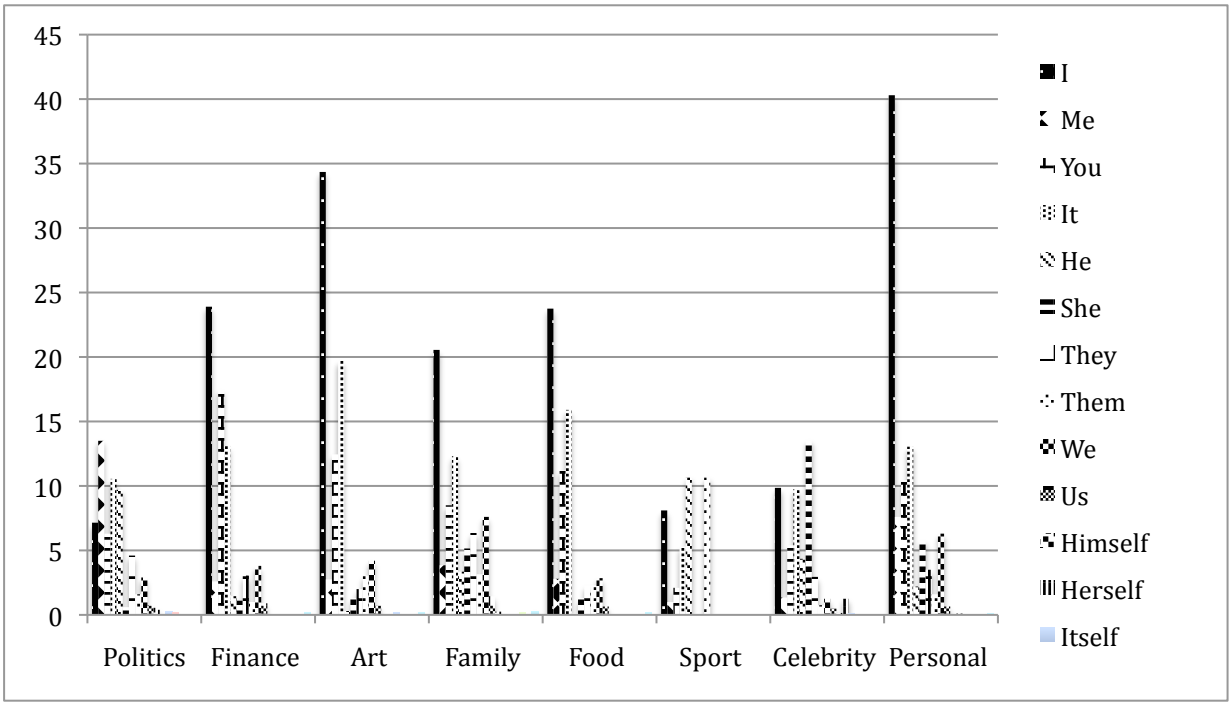

Figure 6. Frequency of each personal pronoun per 1000 words

From the table and figure above it is evident that the personal blog consisted of the highest frequency of personal pronouns. This is not surprising given the personal focus of the Personal blogs. The use of the personal pronoun *I* was by far the most frequent in the Personal blogs and further supports the claim that Personal blogs focus mainly on the blogger's own personal life rather than provide information about things external to the blogger's life.

The family, art and Finance blogs also scored high on personal pronouns. The reason for Family blogs scoring high on personal pronouns is perhaps self-explanatory. The Family blogs discuss issues that include the bloggers own experiences and because the topic of family is to a large extent personal the frequent use of personal pronouns is expected. The Art blogs showed a high frequency of first person pronouns especially *I* due to the bloggers frequent reference to art work completed by them. The Finance blogs also had a high frequency of first person pronouns, which is also due to the bloggers discussing their own life in relation to financial issues or strategies.

The use of third person pronouns was relatively infrequent across the board. However, the celebrity and Sports blogs displayed perhaps the most significant results. The Sports blogs made frequent use of *he* suggesting a dominance of males in the Sports blogs whilst the Celebrity blogs made frequent use of *she* again suggesting that females dominated the Celebrity blogs.

The 'information versus involved production' dimension identified by Biber (1988) is useful in interpreting these results. The more involved the text, the more frequent the use of personal pronouns. However, what is significant here is that the blogs in this study can be seen along a continuum ranging from the most involved to the least involved.

6.5.2. Demonstrative pronouns

Demonstrative pronouns have the function of making something known as well as specifying its distance (Biber et al 1999:347). Demonstrative pronouns have definite meaning and their reference depends on the context shared by addresser and addressee. There are four demonstrative pronouns: *this*, *that*, *those* and *these*. Demonstratives are typically used for spatial or temporal deixis. Demonstratives that point to referents in the near distance or time are *this* and *these* whilst demonstratives that point to referents in the far distance or time are *that* and *those*.

In terms of their distribution across registers, Biber et al. (1999:349) found that demonstratives were rare compared to personal pronouns and that they are more frequent in conversation than in any other register. The most common demonstrative pronoun is *that*. In the academic registers, *this*, *these* and *those* are more frequent than in the other registers due to their use in "marking immediate textual reference" (Biber et al.1999: 349). It must be noted however that because *that* is a flexible word form, which can also function as a relative pronoun, complementizer or degree adverb, all instances of *that* were checked manually in this study.

Table 13. Frequency of demonstratives in each of the blog topics per 1000 words

	Politics	**Finance**	**Art**	**Family**	**Food**	**Sport**	**Celebrity**	**Personal**
Demonstratives	22.05	21.25	21.45	20.5	18	18.7	18.35	21.6

The results show that the Food blogs had the lowest frequency of demonstrative pronouns whilst the Politics blogs had the highest frequency of demonstrative pronouns. This indicates that Political blogs make frequent use of directing the reader's attention to different links and events for example "see this link". This suggests that the Politics blogs need referential cohesion by using demonstrative pronouns. Perhaps the most striking thing however is the similarity between the blog topics. This result also mirrors the result from the Brown corpus where the frequency of demonstrative pronouns was distributed rather evenly across the different types of texts.

6.5.3. Possessive pronouns

Possessive pronouns primarily deal with relationships of possession although their role is not confined just to possession. Biber et al. (1999) state possessive pronouns "are typically used where the head noun is recoverable from the preceding context" (Biber et al.1999: 340). Biber et al. (1999:340) give the following example to clarify:

1. Could be- the same [car] as- **ours**.

In terms of corpus distributions, Biber et al. found that because possessive pronouns are "restricted in their grammatical distribution" (Biber et al 1999:342) since they need a head noun they are rare in comparison with personal pronouns, which occur much more frequently.

They also found that first and second person possessive pronouns are the most frequent forms of possessive pronouns used especially in conversation. In addition they found that masculine forms of the third person pronouns are more frequent than feminine third person forms. Third person pronouns were more frequent in fiction than any other register. Biber et al. also state that the high frequency of possessive pronouns in the conversation and fiction register is due to the high frequency of genitives and ellipsis in those registers, which facilitates the frequent use of possessive pronouns without a head noun.

Table 14: Frequency of possessive pronouns in each of the blog topics

	Politics	**Finance**	**Art**	**Family**	**Food**	**Sport**	**Celebrity**	**Personal**
Possessive pronouns	15.7	20.4	21.55	29.5	14.6	16.5	27	26

The family blog consists of the highest frequency per 1000 words of possessive pronouns followed by the celebrity and Personal blogs. This result is to be expected considering the nature of these blogs. The family and Personal blogs pertain to people and their relationships. The Celebrity blogs consist of gossip surrounding celebrities and their possessions, records, event, ornaments, shows and so forth. Thus the use of possessive pronouns is to be expected as possessive pronouns indicate ownership. In addition, this result indicates the conversational nature of the family, celebrity and Personal blogs since possessive pronouns are more frequent in conversation than in formal writing. This suggests that these three blogs in particular are to a certain extent contextual and informal. Ellegård (1978) also found possessive pronouns to be more frequent in the Popular Fiction texts than any of the other texts. This perhaps indicates the contextual nature of Popular Fiction texts, which can be considered as being on the opposite side of the formality scale to the Science texts. Indeed Ellegård (1978) states " As regards most stylistic features, we find popular texts at one end of the scale, and the scientific ones at the other" (Ellegård 1978: 9).

6.6. Contractions

Biber et al. (1999) identified two types of contractions: verb contractions and not-contractions. Contractions were identified as being one of the main features of speech and occurred frequently in the conversation and fiction registers but very infrequently in academic prose. In addition, contractions after the verb *do* were more frequent than contractions after modal verbs. According to Biber (1988.19) "contractions and first and second person pronouns share a colloquial, informal flavor" because they are used in time-constrained

situations that require immediate response and for this reason they are frequent in informal conversation. Thus the investigation of contractions provides insight into the formality of the different blogs.

Table 15: Contractions frequency per 1000 words in each blog topic

Topic	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
Contractions	15.05	8.85	14.25	9.6	12.5	5.35	12.5	12.3

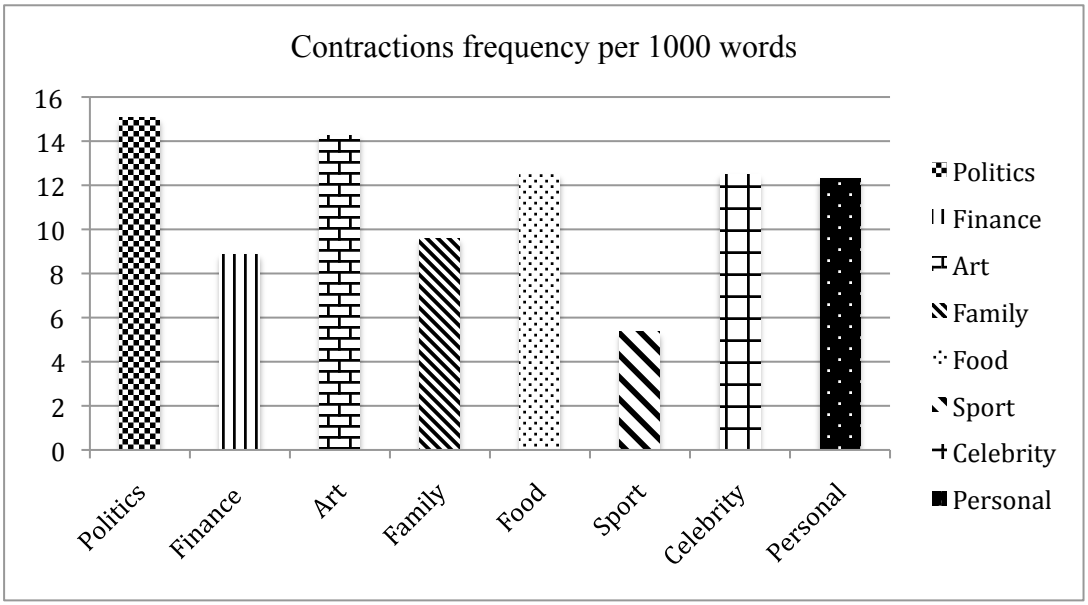

Figure 8. Frequency of contractions per 1000 words.

The contraction frequency results are rather strange and in general do not follow the F-score except for the sports blog result. The sports blog, which had the highest F-score and is thus considered the most 'formal' and the least contextual, also had a very low frequency of contractions per 1000 words. However, for the other blog topics, the contraction frequency did not follow the F-score. This suggest that although the F-score result correlated with the result for the Sports blogs, the F-score cannot be taken as an absolute measure for formality and further supports the argument that the F-score should be taken as a measure of contextuality. The results may indicate that the presumptions concerning the use of contractions may not be relevant for weblogs and that contractions may be used in formal blog topics.

6.7. Hedges

"Hedges are expression[s] of tentativeness and possibility" (Hyland 1996:433 cited in Knight et al. 2013:135). Hedges can operate as face saving devices, to indicate politeness, to tone down the force of an utterance and to indicate vagueness. Knight et al. (2013: 136) state "the frequent use of hedges is often linked to formal rather than informal contexts of communication" and because writing is typically more formal than speech, the frequency of hedging is generally higher in writing. However, it must be stated that other studies have indicated that some types of hedges are more frequent in spoken registers than written registers (e.g. Gries and David 2007). This study examines the frequency of all the hedges used in Knight et al.'s (2013) study as well as the hedges examined in Biber's (1988) analysis of spoken and written registers.

6.7.1. Hedges listed in Knight et al. (2013)

The following table shows the frequency of hedges per 1000 words. The list of hedges is taken from Knight et al. (2013). The most frequent hedges are presented in the table below. There are no totals for this table as it is not the complete list. For the complete table see appendix C.

Table 16. Frequency per 1000 words of hedges listed in Knight et al. (2013) in the different blogs

Hedge	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
Just	2.9	4	3.8	2.1	3.45	1.55	2	2.7
Really	0.85	1.1	0.9	1.6	1.45	0.6	1	1.3
Only	0.75	1.65	0.75	1.3	1.1	2.35	1.2	1.95
Actually	0.4	0.8	0.4	0.25	0.05	0.35	0.5	0.15
Quite	0.15	0.4	0.35	0.2	0.25	0.45	0.1	0.2
Thing	0.7	0.35	0.55	0.2	0.5	0.55	0.6	0.95
Maybe	0.55	0.35	0.3	0.25	0.25	0	0.15	0.7
Probably	0.35	0.9	0.15	0.7	0.45	0.4	0.85	0.15
Of course	0.5	0.2	0.35	0.2	0.2	0.15	0.35	0.1
I think	0.25	1.15	0.85	0.45	0.6	0.25	0.35	0.7
You know	0.4	0.1	0.3	0.05	0.3	0.05	0.35	0.4
Likely	0.45	0.4	0	0.05	0.05	0.45	0	0.05

The Finance blogs followed closely by the Food blogs had the highest frequency of hedges. These topics are the most impersonal and thus considered to fall into the formal category and

consequently a high frequency of hedges is expected. The Personal blogs also scored a relatively high frequency of hedges and this result is also expected because spoken contexts as found by Knight et al. (2013) have a high frequency of hedges. The most frequent hedges were *just*, *only* and *really* for all the blog topics. Again this result is similar to the result found by Knight et al. where the most frequent hedges in the CANELC corpus were the adverb *just* followed by *really* and *only*.

6.7.2.Hedges listed in Biber (1988)

Table 17: frequency per 1000 words of hedges listed in Biber (1988)

Hedge	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
At about	0	0	1.5	0	1.5	0	0	0.5
Something like	0.5	1.5	0.5	0.5	0.5	0	0	0.5
More or less	0	0.5	0	0.5	0	0	0	0
Almost	1	3	2.5	1	2.5	2.5	5.5	3.5
Maybe	5.5	3.5	3	2.5	3	0	1.5	7
Sort of	2	1.5	2	0	2	0.5	0.5	0.5
Kind of	3.5	2	2.5	1	2.5	0.5	0	0
Total	12.5	12	12	5.5	12	3.5	7.5	12

The Politics blogs had the highest frequency of hedges listed in Biber (1988). Biber considers hedges to be informal markers that indicate uncertainty. In this case the Personal blogs displayed a result that is to be expected. However, the Politics blogs displayed a higher than expected frequency of these hedges which are deemed indicative of constrained speech. The Sports blogs displayed a very low frequency of the hedges identified as informal markers, which suggests that the Sports blogs are more towards the formal end of the formality continuum. The result of the Sports blogs appears to be consistent with the F-score. since the Sports blogs had the highest F-score it is expected that they would also have the lowest frequency of informal markers of uncertainty. The Family blogs on the other hand displayed an unexpected result in terms of the frequency of informal markers of uncertainty. The Family blogs had a relatively low F-score so it would be expected that the use of informal features would be high. However, one possible explanation is that the bloggers are sure about the content that they are producing and because the blog is asynchronous and not face-to-face there is less need for face saving devices or use of hedges to soften claims.

6.8. Downtoners

Biber (1988:240) states “downtoners have a general lowering effect on the force of the verb” (Quirk et al.1985: 597-602). Biber notes that they have been characterized by Chafe and Danielewicz (1986) as hedges because of their frequent use to indicate probability in academic prose. The downtoners examined in this study are the same as Biber (1988): *almost*, *barely*, *hardly*, *merely*, *mildly*, *nearly*, *only*, *partially*, *partly*, *practically*, *scarcely*, *slightly*, *somewhat*. Consider now table 18, which shows the frequency of downtoners for each blog topic.

Table 18. Frequency of downtoners in each of the blog topic

Downtowners	**Politics**	**Finance**	**Art**	**Family**	**Food**	**Sport**	**Celebrity**	**Personal**
Almost	0.1	0.3	0.25	0.1	0.3	0.25	0.55	0.35
Barely	0.1	0	0	0	0.2	0	0	0.15
Hardly	0	0	0.1	0	0.05	0	0	0
Merely	0	0	0	0	0	0	0	0
Mildly	0.05	0	0	0	0	0	0	0
Nearly	0.2	0.4	0.05	0.05	0.1	0.1	0.15	0
Only	0.75	1.65	0.75	1.3	1.1	2.35	1.2	1.95
Partially	0	0	0	0	0.05	0	0	0
Partly	0	0	0	0.05	0	0	0	0.05
Practically	0	0.05	0	0	0	0	0.05	0.1
Scarcely	0	0	0	0	0	0	0	0
Slightly	0.05	0.05	0.05	0.25	0.5	0	0	0.05
Somewhat	0	0	0	0.1	0	0.15	0.05	0
Total	1.25	2.45	1.2	1.85	2.3	2.85	2	2.65

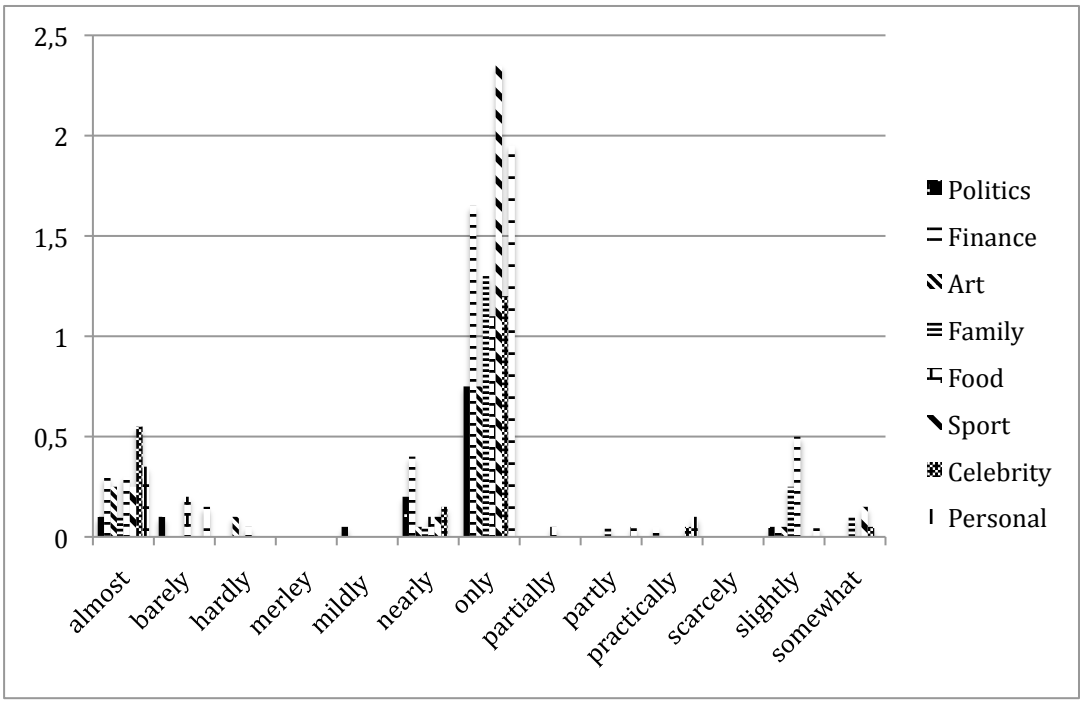

Figure 9. Frequency of downtoners in each of the blog topics.

It is clear from the table and the graph especially that one downtoner is used the most frequently in all the blog topics. The downtoner *only* had a strikingly high frequency in comparison to the other downtoners. The Sports blogs in particular showed the highest frequency of downtoners followed by the Personal blogs. This result indicates the popularity of this downtoning hedge as it was also found to be amongst the most frequent hedges in Knight et al. (2013). Over all however, downtoners excluding *only* occur infrequently in all the blog topics.

6.9. Amplifiers

Amplifiers are the opposite of downtoners, where downtoners lower the effect of the verb, amplifiers "boost the force of the verb" (Quirk et al.1985: 590-7). The amplifiers investigated in this study are: *absolutely*, *altogether*, *completely*, *enormously*, *entirely*, *extremely*, *fully*, *greatly*, *highly*, intensely, perfectly, strongly, *thoroughly*, *totally*, *utterly* and *very*.

Table 19: frequency of amplifiers per 1000 words in each blog topic

Amplifiers	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
Absolutely	0.05	0.1	0.1	0	0.2	0.1	0.15	0.2
Altogether	0	0	0	0	0	0.05	0.05	0.05
Completely	0.15	0.2	0.15	0.1	0.65	0	0.05	0.15
Enormously	0	0	0	0	0	0	0	0
Entirely	0.1	0	0	0.15	0	0.1	0	0
Extremely	0.05	0.35	0	0.05	0.15	0.2	0.15	0.05
Fully	0.15	0.15	0.05	0.15	0.3	0.25	0	0.15
Greatly	0.05	0.1	0	0.05	0	0	0.05	0
Highly	0.05	0	0	0.05	0.15	0.05	0.15	0.15
Intensely	0	0	0	0.05	0	0.05	0.05	0.05
Perfectly	0	0	0.3	0	0.2	0.05	0.05	0
Strongly	0.05	0	0	0	0	0.05	0	0
Thoroughly	0	0	0	0	0.05	0	0	0.05
Totally	0	0	0.15	0.05	0.1	0	0.1	0.4
Utterly	0	0	0	0	0	0	0	0
Very	1.1	1.5	1.3	1.45	1	0.8	1.15	1.15
Total	1.75	2.4	2.05	2.1	2.8	1.7	1.95	2.4

The Food blogs had the highest frequency of amplifiers followed by the Personal blogs and the Finance blogs. The amplifier *very* was the most frequent in all the blog topics. The use of amplifiers can be considered subjective because amplifiers "boost the force of the verb" (Quirk et al.1985: 590-7) and thus they can be used to reflect the attitudes of the bloggers towards certain things. Perhaps the Food blogs having the highest frequency of amplifiers reflects the subjective nature of attitudes towards food and taste. Thus the food bloggers often made use of clusters such as "very tasty" indicating the degree to which the food described tasted well. The frequent use of amplifiers in the Personal blogs is to be expected due to the subjective nature of personal blogs, however the fact that Finance blogs had the same frequency of amplifiers as the Personal blogs was not expected. However a closer inspection of the Finance blogs reveals that the bloggers made frequent use of personal experience and reference to self while discussing financial matters.

6.10. Emphatics

Emphatics indicate involvement and are frequent in informal discourse. Biber (1988:241) states "emphatics simply mark the presence versus the absence of certainty". Emphatics had a positive weight in Biber's (1988) MD analysis indicating an association with involved non-informational production. Emphatics also indicated a personal focus in Grieve et al.'s (2010)

MD analysis. The emphatics investigated in this study are *for sure*, *a lot, such a, real +ADJ, so+ ADJ, DO+V, just, really, most and more.*

Table 20: Frequency of emphatics per 1000 words

Emphatics	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
For sure	0	0	0.05	0.05	0	0.05	0.05	0
A lot	0.55	1.05	0.6	0.4	0.4	0.4	0.6	0.4
Such a	0.6	0	0.25	0.1	0.25	0.15	0.25	0.05
Real+ adj	0.45	0.35	0	0.05	0	0.05	0.25	0.05
So +adj	0.4	0.2	3.65	1.2	1.55	0.2	0.95	1.8
Do + v	0.2	0.15	0.25	0.1	0.1	0	0.05	0.05
Just	2.7	4	3.8	2.1	3.45	1.55	2	2.7
Really	0.7	1.1	0.9	1.6	1.45	0.6	1	1.3
Most	0.6	0.75	0.55	1.35	0.9	1.15	0.8	1.2
More	2.7	3.7	2.7	2.95	2.25	2.35	1.75	1.85
Total	8.9	11.3	12.75	9.9	10.35	6.5	7.7	9.4

The table above shows that the Art blogs had the highest frequency of emphatics overall. A closer inspection of the Art blogs suggests that emphatics play an important role in the description of the items in the Art blogs. Clusters such as “so pretty”, “so girly”, “really love” and “really like” are frequent in the Art blogs. The Finance blogs also had a rather high frequency of emphatics especially the emphatic *just*. The Sports blogs on the other hand displayed a very low frequency of emphatics. This low score again confirms the more formal and impersonal nature of the Sports blogs in this study since emphatics typically indicate personal focus and involvement.

6.11. Discourse particles

Biber et al. (1999:1086) describe discourse particles as inserts that occur either at the start of an utterance or at a turn of an utterance. They have the function of signaling a change in the direction of an utterance or indicating interaction between interlocutors.
In addition, discourse particles monitor the flow of information in conversation and Biber (1988: 241) states that they are “rare outside of conversational genres”. The discourse particles investigated in this study are the same as the discourse particles investigated in Biber (1988): *well*, *now*, *anyway*, *anyhow*.

Table 21. Frequency of discourse particles per 1000 words

Discourse particle	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
Well	1.05	1.3	1.05	0.5	1.3	0.9	0.8	1.3
Now	1.85	1.7	1.65	1.2	1.1	0.85	1.3	2.15
Anyway	0.15	0.15	0.1	0.15	0.15	0.1	0.05	0.05
Anyhow	0	0	0	0	0	0	0	0.05
Anyways	0	0	0	0	0	0.05	0.05	0.25
Total	3.05	3.15	2.8	1.85	2.55	1.9	2.2	3.8

From the table above it is evident that the Personal blogs used discourse particles the most. This indicates that the Personal blogs are conversation-like in this regard. The Sports blogs had a low frequency of discourse particles and again this supports the F-score result that gave the Sports blogs the highest F-score and thus the most 'formal'. However, the Family blogs surprisingly had the lowest frequency of discourse particles. This suggests that the Family blogs are structured in a way where discourse particles are not needed to change topic or control the flow of information. This is perhaps due to each blog post focusing on one particular issue rather than having a number of issues in one blog post. Due to the scope of this study, this finding could not be fully investigated. Further research however could examine the structure of the Family blogs and investigate how the flow of information is structured. Nonetheless, overall the blogs in this study had a low frequency of discourse particles and this suggests that the blogs are more formal than conversation.

6.12. Private verbs

Quirk et al. divides factual verbs into two parts: public and private. A private verb "expresses intellectual states such as belief and intellectual acts such as discovery" (Quirk et al. 1985:1181) such as *accept*, *believe* and *check*. These types of verbs are private because the process cannot be observed (Quirk et al. 1985:1181). The private verbs investigated in this study are all the private verbs mentioned in Quirk et al. (1985:1181). In addition, all forms of the private verbs were searched for and not just the base form of the verbs. The most frequent private verbs are presented in the table below. There is no total for this table as it is not the complete list. For the complete list of verbs with totals see appendix D.

Table 22. Frequency of private verbs per 1000 words in each blog topic

Private verb	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
Accept	0.05	0.05	0	0	0.05	0.05	0	0.05
Believe	0.75	0.1	0.15	0.3	0.05	0.3	0.4	0.25
Consider	0.25	0.15	0.15	0.2	0.15	0.15	0.1	0.05
Decide	0.05	0.1	0.4	0.35	0.05	0.05	0.45	0.55
Dream	0.05	0.15	0.1	0.2	0.4	0	0.55	0.4
Feel	0.25	0.65	0.4	1.7	0.5	0.6	0.4	0.6
Find	0.55	1	0.8	1.5	0.65	0.55	0.3	0.65
Guess	0.15	0.2	0.25	0.2	0.1	0.1	0.55	0.05
Hear	0.75	0.2	0.5	0.35	0.2	0.05	0.05	0.75
Hope	0.35	0.2	1.55	0.35	0.2	0.3	0.65	0.5
Know	2.4	1.65	1.8	1.6	1.65	0.85	1.3	2.5
Mean	0.25	1.15	0.4	0.9	0.5	0.35	0.45	0.55
Remember	0.1	0.3	0.05	0.7	0.15	0.15	0.3	0.2
See	0.85	1.4	1.7	1.1	0.75	0.85	0.8	2.45
Show	1.2	0.65	0.6	0.05	0.05	0.5	0.55	0.6
Think	1.25	2.7	1.7	1.4	1.1	1.1	1.95	2.65
Understand	0.2	0.2	0	0.95	0	0	0	0.3

The Family blogs, closely followed by the Personal blogs, scored the highest in the frequency of private verbs. The private verbs indicate an involved approach and thus it is expected that the Personal blogs in particular would have a high frequency of private verbs. The most frequent private verbs were *see*, *show*, and *think*. These verbs are frequently used in everyday conversation hence their high frequency in the blogs in comparison to the other private verbs indicates that the blogs are conversational to some degree. On the other hand, the Food blogs had the lowest frequency of private verbs followed by the Sports blogs. This suggests that these two blogs are less involved and more formal in comparison to the Family and Personal blogs.

6.13. Modal and semi modal verbs

This section discusses modal verbs. Biber et al. (1999) found that modals differ greatly in their distribution across registers. Thus modals and semi-modals were more frequent in conversation and the least frequent in academic prose and news (Biber at el.1999: 486). Biber et al. also found that semi modals in particular displayed striking results and were five times

more frequent in conversation than written texts. This study will thus examine the frequency of modals to ascertain which blog topics are closer to conversation and thus by association informal in terms of modal frequency and which topics are closer to the more formal registers. The blogs in this study were searched for the frequency of both modal and semi modal verbs.

Table 23. Frequency of modal verbs per 1000 words

Modal	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
Will	2.95	4.85	3	2.95	1.9	3.85	3.05	2.75
Would	2.3	3.45	1.3	3.85	1.25	3.95	1.55	1.8
Can	1.9	5.9	3.1	4.8	3.3	1.5	2.35	2.6
Could	0.85	1.35	0.7	2	1	1.3	0.6	0.75
May	0.95	2.7	0.45	0.85	0.2	0.6	0.75	0.7
Should	0.8	1.25	0.2	0.8	0.65	1.75	0.45	0.85
Must	0.25	0.25	0.35	0.45	0.25	0.35	0.25	0.35
Might	0.65	0.45	0.15	0.5	0.6	0.9	0.4	0.35
Shall	0	0	0	0	0	0.15	0	0.05
Total	10.65	20.2	9.25	16.2	9.15	14.35	9.4	10.2

In terms of modal verb use, the Finance blogs had the highest frequency. This is rather surprising considering the formal nature of financial texts. However bearing in mind that the financial blogs examined in this study displayed a highly involved writing style with a high frequency of personal pronouns, this result then becomes understandable especially since Biber et al. (1999) found that modal verbs and semi modals were frequent in conversation and less frequent in academic prose.

The Sports blogs, which had a high F-score and thus was deemed to be the least contextual, also had a high frequency of modal verbs. This result supports Herring et al.'s (2004) claim that blogs are a hybrid genre. The Sports blogs display impersonal information on the one hand but also incorporate features common in spoken contexts.

The Personal blogs had a lower frequency of modal verbs than the Finance, Family and Politics blogs. This result is perhaps also unexpected as Personal blogs are deemed to be the most contextual and thus the least formal and closest to conversation and thus the frequency of modal verbs is expected to be high. However, again the notion of hybrid genre can provide an explanation for this.

Semi modal verbs

Table 24. Frequency of semi modal verbs per 1000 words in each blog topic

Semi modals	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
Had better	0.1	0	0	0	0	0	0	0
Have to	0.55	1.2	0.5	0.7	0.45	0.5	0.1	0.85
Have got to	0	0	0.05	0	0	0	0.05	0
Gotta	0	0	0	0	0	0	0	0
Supposed to	0.2	0.15	0	0	0.05	0	0.2	0
Going to	0.6	0.75	0.7	0.65	0.45	0.65	0.9	1.15
Gonna	0.2	0	0	0	0.05	0	0.05	0
ve gotta	0	0	0	0	0	0	0	0
m gonna	0	0	0	0	0	0	0	0
Total	1.65	2.1	1.25	1.35	1	1.15	1.3	2

The frequency of semi modals is low in comparison with the modal verbs. This perhaps is due to semi modals being used more in conversation than in written contexts. In addition, Biber et al. (1999) found that because semi modals are considered to be recent, they are more established in speech than in writing. However it must be noted that the Personal blogs and the Finance blogs had the highest frequency of semi-modal verbs suggesting a similarity with conversation in this regard.

6.14. Expletives

Expletives refer to "taboo expressions (swear words) or semi- taboo expressions used as exclamations, especially in reaction to some strongly negative experience" Biber et al. (1999: 1094). Expletives are thus most frequent in casual conversation and rare in more formal prose. The list of expletives examined is from Biber et al. (1999:1098).

Table 25. Frequency of expletives per 1000 words

Expletives	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
God	0.55	0	0.1	0.2	0	0	0.1	0.35
My god	0	0	0	0	0	0	0	0
Christ	0	0	0	0	0	0	0	0
Damn	0.1	0.05	0	0.15	0.05	0.2	0.4	0.4
Hell	0.2	0.05	0	0.1		0	0.1	0.25
Shit	0.15	0	0	0	0	0.05	0.05	0.1
Fuck	0	0	0	0	0	0.25	0.1	0.2
Good lord	0	0	0	0	0	0	0	0.05
Heck	0	0.1	0	0	0.05	0	0	0.05
Bullshit	0.05	0.05	0	0	0	0	0	0
Darn	0.05	0	0	0	0	0	0	0
Goodness	0	0	0.05	0	0	0	0	0
Total	1.1	0.25	0.15	0.45	0.1	0.5	0.75	1.4

From the table above it is evident that the frequency of expletives in the blogs is quite low. However, the Personal blogs had the highest frequency of expletives. This result indicates that the Personal blogs are quite informal and share this feature with casual conversation, which is the register that normally has the most frequent use of expletives.

6.15. Word and sentence length

According to Biber (1988:104) "longer words also convey more specific, specialized meanings than shorter words". In addition, Biber cites Zipf (1949) who found that shorter words are more frequently used and thus consequently have a more general meaning. This study examines the mean word and sentence length for each blog topic to ascertain if there is a difference between the blog topics in terms of sentence or word length. *Wordsmith* concordance program calculated the mean word length and mean sentence length in characters. These results are presented in the table below.

Table 26. Mean word length and mean sentence length for each of the blog topics.

	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal	Brown corpus
Mean word length	4.77	4.3	4.26	4.32	4.41	4.45	4.55	4.21	4.38
Mean sentence length	17.32	16.85	13.61	18.23	15.93	21.83	17.88	13.19	18.36

The results do not indicate a great variation between the mean word lengths of the different blog topics. Nonetheless the Politics blogs have the longest mean word length of 4.77 characters whilst the Personal blogs have the shortest mean word length of 4.21. The shorter mean word length in the Personal blogs again indicates the conversational nature of the words used in this type of blog. According to Heylighen and Dewaele (1999) informal discourse uses shorter words that are more direct and this appears to be the case in the Personal blogs. The Personal blogs also had the shortest sentence length whilst the Sports blogs had the longest mean sentence length. These findings again confirm the findings of the F-score. The sports blogs were the least contextual texts whilst the Personal blogs were the most contextual.

6.16. F-scores for individual blogs

The following sections examine each blog in detail providing an F-score for each of the four blogs from each topic in order to ascertain whether the F-score is consistent for each blog or whether one blog skews the overall blog topic results. For each blog topic, one example is given of the most contextual blog and one of the least contextual blog. All instances of personal pronouns, which are considered to be the most obvious indicators of involved contextual style, are highlighted in bold.

6.16.1 Political blogs

Consider now the table below, which shows the F-scores for each of the four Political blogs used in this study.

Table 27. F-score of the individual Politics blogs

	Politics blog 1	Politics blog 2	Politics blog 3	Politics blog 4
F-score	65.27	68.86	63.33	66.31

The Politics blogs all scored over 60 in the F-score. This indicates that all the Politics blogs are focused on imparting external information and are less contextual. The first three blogs were written by males whilst the forth blog was written by a female.

6.16.2 Finance blogs

Table 28. F-score s of the individual Finance blogs

	Finance blog 1	**Finance blog 2**	**Finance blog 3**	**Finance blog 4**
F-score	55.70	61.81	53.26	68.86

The Finance blogs showed some variation in terms of the F-score with Finance blog one having a low F-score of 55.70 and Finance blog 4 having and F-score of 68.86. Finance blog 1 was written by a male engineer who quit his engineering career to become a stay at home dad. The blog is based around the blogger's personal experiences in the financial world and this perhaps explains the low F-score for this blog. Below is an example from Financial blog 1.

Example 1

> One of **my** New Year Resolutions is the No New Clothes Challenge for 2013. **I** figured it shouldn't be that hard because **I** have plenty of clothes and shoes. Since **I** don't work outside the home anymore, **I** don't really need to be that presentable and it will be fine if **I** just wear some clothes down.
>
> Apparel spending is one of the smaller expenses in consumer expenditures so even if **you** cut **your** budget here, it won't affect **your** expenses that much. According to the US Census Bureau's data, apparel is only about 3.5% of an average family's annual expenditure.

The above extract exemplifies the subjectivity incorporated in this blog and the frequent use of self-reference to exemplify how to reduce spending. In 101 words 9 instances of personal pronouns are found and 6 of them focus on the blogger.

Finance blog 4 on the other hand had the highest F- score out of al the Finance blogs. Finance blog 4 was written by a male college graduate and it is written in a less personal style than Finance blog 1. The extract below from Finance blog 4 exemplifies this.

Example 2

> Here is **my** May 2013 update of the trailing total returns for selected major asset classes. Passive ETFs are used to represent major asset classes, as **they** represent actual investments that folks can buy and sell. Return data was taken after market close at the end of April 2013.
>
> **I**'m trying out a new chart format, in the hopes of easier visual comparisons. Below is a chart of the all the trailing returns for 1-month, 1-year, 5-year, and 10-year periods.
>
> If **you** focus on the blue and red bars, **you** can see that in the short-term the stock markets around the world have been on quite a tear.

The extract above shows a more formal style than the previous extract hence the higher F-score for this blog. Notice the focus is not mainly on the author thus use of personal pronoun *I* only occurs once. Instead the focus is more on the reader as shown by the use of *you.* In 108 words 5 instances of personal pronouns are found and only two of them focus on the blogger.

6.16.3. Art blogs

Table 29. F-scores of the individual Art blogs

	Art blog 1	Art blog 2	Art blog 3	Art blog 4
F-score	59.46	59.13	65.74	55.07

The Art blogs also showed some variation in terms of F-score result. Art blog 4 had the lowest F-score whilst art blog 3 had the highest F-score. All the Art blogs are written by women. The blogger of art blog 3, which is the most formal out of all the Art blogs, also sells art supplies from her blog. The extract below provides an example from art blog 3.

Example 3

> Today **you're** in store for super inspiring projects from the industry's leading designers, and **we're** also giving away a Taylored Expressions/Authentique prize pack to one lucky winner! The winner will be chosen from among all of the comments left on participating blogs during **our** two day hop. **You** have until Monday, April 15th at 11:59 PM PST to leave **your** comments.

Notice that instances of first person pronouns are low. There are no instances of *I* rather the text focuses on the reader hence the use of you a number of times. Consider now the following extract is from art blog 4 which had the lowest F-score.

Example 4

> **My** daughter turned 7 yesterday. **I** can hardly believe it! Since this was the year of the HUGE pillow at **our** house, **I** just *had* to make one for **her**, too. These things take longer to make than **you** would think (about a week!) but **they** are totally worth it.
>
> **My** girl just LOVES purple. (Figures, doesn't it? It sure isn't **my** favorite.)

The above extract shows how the blogger make frequent reference to self and family notice the frequent use of *I*, *me* and *my*. The text is more involved than informational in comparison to the extract from art blog 3, which is more informational rather than involved which has probably contributed to the low F-score for art blog 4.

6.16.4. Family blogs

Table 30. F-score of the individual Family blogs

	Family blog 1	Family blog 2	Family blog 3	Family blog 4
F-score	57.73	56.59	61.39	54.15

The Family blogs showed slight variation in terms of the F-scores of the different blogs. Family blog 4 had the lowest F-score whilst family blog 3 had the highest F-score. The first three Family blogs were written by women whilst family blog 4 was written by a male. Family blog 3 which had a high F-score was written by a female blogger who is also an educational consultant, a qualified teacher and has a Master's degree. Family blog 4 was written by a stay at home dad who wrote his blog to demonstrate what great dads do to demonstrate their love. The following extract is from family blog 3.

Example 5

> iTooch Elementary School is an educational app for the iPad and iPhone designed for children in 3rd, 4th and 5th Grade. iTooch is developed by a team of teachers, experts, parents and video game designers to help reinforce concepts and material learned at school.
>
> Most effective learning takes place in short focused sessions, the app is designed to do just that, by encouraging **your** child to learn whilst having fun.

It is clear that the extract above is highly informational as indicated by the absence of first person personal pronouns. The focus is on imparting information rather than an involved approach. The text makes no reference to the blogger at all. When compared to the extract from family blog 4 the distinction becomes clear.

Example 6

> With four kids between the ages of 11 and 20, **we're** living through the cell phone challenge. **We've** asked ourselves several times the questions that most parents do: When is the right time to give **my** child a cell phone? What type of phone and plan? How do **we** monitor the phone's use? What are the consequences of cell phone abuse? Should the child help pay for the phone and/or plan?

Notice that that example 6 makes frequent use of *we* in the extract above, which refers to the blogger and his wife.

6.16.5. Food blogs

Table 31. F-score of the individual Food blogs

	Food blog 1	Food blog 2	Food blog 3	Food blog 4
F-score	61.20	60.97	65.23	65.59

The Food blogs all scored above 60 on the F-score. This suggests that Food blogs are generally formal and less contextual in nature. However, there is some slight variation between the different food blogs and Food blog 2 had the lowest F-score whilst food blog 4 had the highest F-score. All the bloggers of the Food blogs were women. Both the bloggers of food blog 2 and Food blog 4 have experience writing on food topics. The blogger of Food blog 2 has written a cookbook and the blogger of Food blog 4 is a food writer and restaurant reviewer. However they have different styles of writing with Food blogger 2 writing in a more involved approach than food blogger 4. The extracts below exemplify this further. The following is an extract from Food blog 2.

Example 7

> Who's ready for spring produce? Picture **me** raising my hand really high, like a nerd in the front row of math class. While **I** love root vegetables of all kinds, **I**'m ready for some nice crunchy asparagus! **I** eat a metric ton of asparagus all through the spring and summer - mostly grilled, but **I**'ll take it anyway **I** can get it.
>
> The word *primavera* means "spring" in Italian, so even though **we've** had a rainy, cold spring here in Michigan, **I** decided to brighten up my dinner with this light and fresh pasta dish. This is another one of those meals that instead of serving a bowl of pasta with some vegetables on the side, **I** just put it *all* in the pot and stir**. I** like to do that. So easy.
>
> The great thing about this dish is that you can use any vegetables **you** want. It's also a great way to get **your** family to eat their veggies while still allowing them to think **they're** just eating a big bowl of pasta for dinner.

It is evident that the extract above is written in an involved style with frequent use of personal pronouns especially the first person *I*. Consider now an example from Food blog 4 and notice how the frequency of personal pronouns differs.

Example 8

> Most of the Nespresso grand crus capsules are blends of beans from different locations to achieve a specific flavor profile or beans from one location that best represent the flavors of that region. The honey notes of the Brazilian Dulsao da Brazil are incredible in a late afternoon latte. Nespresso also produces limited edition capsules to feature a certain harvest or experience. Trieste and Napoli are the current limited edition coffees available and the Trieste is intense with heavy caramel overtones. **I** adore it. The experience of choosing a jewel colored Nespresso capsule is similar to picking a nice bottle of wine. Each capsule is rated by intensity and flavor notes. Whether you like your coffee dark and intense like me, or prefer a lighter cup to start your day, **you'll** have plenty to choose from in the starter capsule box. There is even a good selection of decaffeinated capsules.

The extract above is not as involved as the previous extract and consists of a lower frequency of personal pronouns.

6.16.6. Sport blogs

Table 32. F-score of the individual Sports blogs

	Sports blog 1	Sports blog 2	Sports blog 3	Sports blog 4
F-score	66.76	68.80	68.82	62.71

All of the Sports blogs are written by males and there does not appear to be a big difference between the Sports blogs F-scores. They all appear to be more formal and less contextual and all had a score over 62. The blog with the lowest F-score was blog 4 whilst the blog with the highest was blog 3. The author of Sports blog 3 is an attorney who likes to blog for fun. The extract below is from sports blog 3.

Example 9

> From his controlled-rage of a swing to **his** aggressiveness on the base paths, Bryce Harper plays the game of baseball with an intensity that most fans normally associate with football . . . or mixed martial arts. For this reason, Harper is often described as "gritty" and "hard-nosed" and any number of other adjectives usually reserved for the scrappy middle-infielder-type who always has **his** uniform dirty. Unlike the light-hitting middle-infielder who is hailed for doing the "little things" to help a team win (i.e., Ryan Theriot or David Eckstein), however, Harper is also unworldly talented, has the potential to be a perennial MVP candidate, and is universally loved by baseball traditionalists and Sabermetricians alike. Harper is the rare star athlete who has the mentality of the last player on a roster: "I'm going to play every single game like it's **my** last."

The extract above is descriptive and informative rather than involved and interactive notice the low frequency of personal pronouns in this extract. The extract below shows a more involved example from Sports blog 4 with a higher frequency of personal pronouns.

Example 10

> In a couple of weeks **I**'ll give my son **his** big 7th birthday present: A Russell Wilson jersey. **He's** shown interest in the Seahawks for a while, and after taking **him** to a Columbus Blue Jackets game last week, I know **he's** got the "spazztacular sports fan" gene. The kid was LEADING "Let's Go Jackets" chants all night long, so **I'**m pretty sure that once **I** take him to a Seahawks game (perhaps at Indianapolis this fall), **he'll** become a Twelve for life (and **my** 3-year-old daughter already likes the Seahawks too, but **I'**ve got more time to work on **her**).

Notice the high frequency of personal pronouns, which make this text interactive and involved as opposed to informational and descriptive.

6.16.7. Celebrity blogs

Table 33. F-score of the individual Celebrity blogs

	Celebrity blog 1	Celebrity blog 2	Celebrity blog 3	Celebrity blog 4
F- score	63.99	65.00	58.94	66.72

The Celebrity blogs show some variation in the F-scores of the different blogs. All the Celebrity blogs were written by women. Blog 4 had the highest F-score whilst blog 3 had the lowest. The following extract is from blog 3.

Example 11

> Lindsay Lohan will make a mockery of rehab this weekend. But before **she** leaves for treatment **she's** cramming in as many of **her** favourite activities as possible, including bar-hopping in New York City and illegal driving. This time, the 26-year-old's Porsche Panamera has been towed because **you're** somehow not allowed to park wherever the hell **you** want just like **you're** not allowed to run over anyone **you** want. According to TMZ.com, Lohan parked illegally in Brooklyn and got towed on Saturday. Even though **she** never found the coins to pay **her** old lawyer, **her** storage fees, and the million other tabs **her** Johns failed to pick up **I** guess there was enough money on **her** nightstand to get the car out of the impound lot.

Notice the frequency of personal pronouns especially the third person *she*. This extract also demonstrates the use of the negative key word *she* in context. Consider now the following extract is from Celebrity blog 4 and notice the difference in pronoun use.

Example 12

> Chris Brown released **his** latest single and music video 'Fine China' on April 1 – the first single off **his** upcoming album titled *'X.'*
>
> According to Chris Brown the song is a tribute to Michael Jackson and one look at the video will confirm it. Rolling Stone accurately call the debut single 'Fine China' a:
>
> The Twittersphere seem to be reacting very positively to the single – many loving the infectious beats, MJ'esque dancing, singing and wholesome feel.
>
> The music video depicts Brown with his forbidden 'Fine China' a pretty girl who has an overbearing father – and the conflict that ensues, naturally on the dance floor (go the dance fight!). Chris Brown is dressed in suspenders and a driving hat and dances his tail off the entire video – bar the suspense filled fight scenes. The lengthy video is very reminiscent of Michael Jackson's the attractive object of affection, and the build up to conflict.

Notice the less frequent use of personal pronouns and the more frequent use of nouns in the extract above, which have probably contributed to the higher F-score.

6.16.8 Personal blogs

Table 34. F-score of the individual Personal blogs

	Personal blog 1	Personal blog 2	Personal blog 3	Personal blog 4
F-score	61.02	45.11	60.09	54.16

The Personal blogs showed a rather wide variation between the different Personal blogs. The lowest F-score was for blog 2 and the highest was for blog 3. The Personal blogs are balanced in terms of bloggers gender. Two Personal blogs were written by males and two were written by females. Blog 2 was written by a female whilst blog 3 was written by a male. Blog 2 is written in a very involved conversational manner whilst blog 3 is written in a more informative way. In blog 2 the focus of the blog is the blogger and her life and her feelings. The extract below from blog 2 exemplifies this.

Example 13

> **I** have been at home and refused to leave for 4 days now. **I** am just not in the mood to deal with anyone now. **My** family is trying to be supportive and mom keeps telling **me** that **I** am not the first one. **I** know. **I** just refuse to be insulted. If **I** was **my** fault, then **I** would understand and let things be. But it is not. That is what is driving **me** crazy.

The example above shows that the focus of the blog is the bloggers feelings. She is upset about an event that happened and is discussing how she feels. Notice the frequent use of the first person pronoun *I*. Blog 3 on the other hand is less involved and more informational. The extract below shows this.

Example 14

> Queenstown is a outdoors enthusiasts adventure playground. A mecca for mountaineering, sailing, trekking (or tramping as the locals call it), skiing in winter, rugby, cycling, rock climbing, canyon swinging, bungee jumping, skydiving, horse riding and every other outdoor activity you could possibly think of. Even Ray Mears couldn't get bored here. The Queenstown air was filled with adrenaline, and **I**'m happy to say it was filling **my** lungs.

Notice the difference between example 13 and example 14. Example 14 is more informative. It displays a lower frequency of personal pronouns and this has probably contributed to the higher F-score for example 14.

7. Discussion

What this overview clearly shows is that there is great variety in the language of blogs and that there is a continuum of linguistic variation among the different blog topics.

One of the main conclusions of this study is that the F- score provides information about the contextuality and involvement of texts. The texts that scored high on the F-score were those which were the most contextual and the most involved. In terms of formality in the traditional sense, the F- score does not necessarily provide a measure of formality. This point is confirmed by Nowson (2006) who states that through discussions with Dewaele, one of the founders of the F-score, it became apparent that the F-score is more of a measure of contextuality rather than formality in the traditional sense. Nonetheless in many cases high contextuality does mean low formality, although this cannot always be guaranteed.

The results show that the F-score is a good measure for contextuality of texts and although it can provide indications of text formality it can not be taken as a measure of formality in the traditional sense as some of the features that indicate informality such as contractions were present in the texts that scored high on the F-score. In addition, nouns proved to be rather problematic especially in topics such as sports where players' names, team names and sports venues were mentioned in a text. It would be a mistake to assume that just because a text had a high frequency of nouns it is more formal. Celebrity blogs also provide a good example where proper nouns in the form of celebrities' names increased the F-score of the blogs when without the proper nouns the F-score would have probably been much lower.

Nonetheless a closer inspection of the blogs shows that the F-score provides an indication of the 'involved versus informational production' dimension identified by Biber (1988). The topics that scored low on the F-score such as the Personal blogs and the Family blogs were more involved whilst the topics that scored high on the F-score such as the Sports blogs and the Political blogs were more informational in nature. This shows that although the F-score cannot be taken as a measure of formality per se, it does provide a good basis for further investigation and it does provide distinctions between text topics. The F-score is also useful for showing a continuum of contextuality amongst the blog topics. From this study it is evident that the Personal blogs are the most contextual whilst the Sports blogs are the least contextual. The results show that the blogs can be placed upon a continuum along the two dimensions *involved* versus *informational* dimension identified by Biber (1988) and *informational versus personal* dimension identified by Daems et al. (2013). The results suggest that the Sports blogs are the most informational and thus have the highest F-score- the Personal blogs are the least informational and thus have the lowest F-score. The results of this investigation also seem to confirm that the two main distinctions in blogs are the personal focus and the informational focus.

The F-score could be used more as a measure of how informational or involved a text is rather than as a measure of formality per se. Nonetheless the F-score does, in many cases, indicate which texts are more formal than others. Upon further inspection, it is clear that texts that are assimilated with informational focus are more formal than texts that are more involved such as the Personal blogs. The results of the F-score thus present the blog topics on a scale from the most informational to the least informational and most involved.

The results indicate that there is a continuum of formality. If we turn to the Finance blogs for instance we can see that despite finance being a topic normally viewed as formal and impersonal, the bloggers discuss financial issues in their own lives so for example we see one blogger discussing a financial experiment he undertook to live on one dollar a month or another blogger discussing whether children should financially help their parents. This personal input in what is otherwise a professional topic adds a personal dimension.

Turning to the Personal blogs it is evident that therein lies the biggest distinction between all the blog topics. The Personal blogs are clearly distinct from all the other blogs in terms of word class frequency such as pronouns and verbs as well as personal input. The active voice in the Personal blogs and the clearly involved approach show that the personal versus thematic dichotomy is clearly the most obvious in classifying blogs. However there also appears to be a continuum in the blog topics. It is perhaps useful to view blogs as conforming to a scale from the most impersonal to the most personal. In addition, in terms of the F-score, which as stated above could be considered as a score of contextuality, it is evident that the more personal information included the lower the F-score.

The results also indicate an idiosyncratic dimension to blogs. The fact that Finance blogs scored lower than the sports, politics and Food blogs indicates that even if the topic is formal and external if the blogger relates the topic to their life and adds a personal dimension as is the case for the Finance blogs, the F-score will be lower and contextuality will be increased. Thus the level of personal involvement rather than blog topic had the greatest effect on the F-score.

The Sports blogs, which had the highest F-score, focused on the external events of the sports rather than personal views on the sport. The sports blog which had a lowest F-score included information about the blogger's first experience at a game and the blogger's son's first experience. It is thus evident therefore that personal involvement plays a big role in the F-score.

To conclude: The F- score is a good indicator of the involved versus informational dimension of texts. Texts which were more informational scored higher on the F- score than

texts that were more involved. For this reason, the F-score is a useful and informative measure of contextuality and information versus involvement of a text rather than the formality of a text in the traditional sense. This is because the texts used in this study, for example the Sports blogs, focused on a high level of imparting information and thus scored high on the F-score. However, to conclude that Sports blogs are formal in the traditional sense would be a mistake. The blogs contained personal pronouns, expletives, irregular spelling and contractions all of which would not occur in formal language in the traditional sense of the word. In addition, this study shows that Herring et al.'s (2004) conclusion that blogs constitute a hybrid genre is relevant. The blogs in this study exemplify this clearly by incorporating various linguistic features belonging to both the formal and informal registers into one blog.

Although this study did not control for gender, the authors of the Sports blogs were all male. On the other hand the Food bloggers were all female. The Family blogs had one male blogger and 3 female bloggers whilst the Personal blogs were balanced with two male and two female bloggers. The celebrity and Art blogs were all written by female bloggers. The Finance blogs were written by three male bloggers and one blogger who used a pseudonym and did not indicate gender. Finally the Politics blogs were written by three male bloggers and one female blogger. Further investigation could examine the effect of author gender or personality on the language of blogs and investigate the relationship between the blog topic and blogger gender.

One of the main limitations of this study is the inaccuracy of the Stanford tagger on irregular texts and unknown words. A future study would improve on this by training the tagger on blog texts in order to gain higher tagging accuracy, but this requires programming knowledge, and was beyond the scope of this study. A future study could also carry out a factor analysis on linguistic variation among the different blog topics to be able to more accurately determine the formality and contextuality of each topic and the influence of genre on blog language.

Future research could investigate the effect of personality on blog contextuality. It has been suggested (e.g. by Nowson et al. (2005), Nowson (2006)) that author personality types influence the contextuality/formality of blogs. More research in this field would increase our understanding of the effect of personality on language production and blog writing in particular.

A more comprehensive study could use more texts. It is difficult to know what the trend will be if a larger number of texts were used. Would the Sports blogs still be the least

contextual? Future research could in addition to using more texts, investigate other variables such as age, gender, socioeconomic status, ethnicity, educational background etc. Other linguistic variables that could prove to be insightful, but have not been investigated in this study due to scope, include the use of emoticons, abbreviations and slang in blogs. Investigation of these variables could provide further assistance in blog classification. A larger study with more texts would help ascertain whether blog writing is influenced only by individual style or whether blog topic really does influence the language of the blogs. This study could thus be used as a platform from which future research on the language of blogs can be developed.

References

Baron, N. (1998). Writing in the age of email: The impact of ideology versus technology. Visible Language 32(1): 35–53. In Knight, D., Adolphs,S., Carter, R. (2013). Formality in Digital Discourse: A study of Hedging in CANELC. *Yearbook of Corpus Lingusitics and Pragmatics 2013: New Domains and Methodologies.* Springer, 131-152.

Baron, N. S. (2008). *Always On: Language in an Online and Mobile World.* Oxford: Oxford University Press.

Biber, D. (1988). *Variation across speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., Johansson,S., Leech,G., Conrad,S.,Finegan, E. 1999. *Longman Grammar of Spoken and Written English.* China: Pearson Education Limited

Biber, D., Conrad, S., Reppen, R., Byrd, P., and Helt, M. (2002). Speaking and Writing in the University: A Multi-dimensional Comparison. *TESOL Quarterly* 36(1), 9-48.

Blood, R. (2002). Introduction. In *We've got blog*, ed. J. Rodsvillz. Perseus Publishing: USA.

Chafe, W and Danielewicz, J. (1986) Properties of Spoken and Written Language in Comprehending Oral and Written Language. New York: Academic Press. In Biber, D (1988)Variation Across Speech and Writing. Cambridge University Press.

Blog. (2013). In Merriam websiter dictionary. Retrieved May 10, 2013, from http://www.merriam-webster.com/dictionary/blog?show=0&t=1375458433

Coffin, C., Jane, Curry, M.J., Goodman, S., Hewings, A., Lillis, T.M., and Swann, J. (2003). *Teaching Academic English: A Tool Kit For Higher Education*. London: Routledge.

Crystal, D. (2001). *Language and the Internet*. Cambridge: Cambridge University Press.

Crystal, D. 2011 *Internet linguistics*: A student Guide. Oxon: Routledge.

Daems, J., Speelman, D., and Ruette, T. (2013). "Register Analysis in Blogs: Coloration between Professional Sector and Functional Dimensions". *Leuven Working Papers in Linguistics* 2 (1), 1-27.

Ellegård, A. (1978). *The Syntactic Structure of English Texts*. Gothenburg: Acta Universitatis Gothoburgensis.

Flowerdew, J. (2013). *Discourse in English Language Education*. Abingdon: Routledge.

Gries, S.Th., and C.V. David. (2007). "This is kind of/sort of interesting: Variation in hedging in English". In Knight, D., Adolphs,S., Carter, R. 2013. "Formality in Digital Discourse: A study of Hedging in CANELC". *Yearbook of Corpus Lingusitics and Pragmatics 2013: New Domains and Methodologies.* Springer, 131-152.

Grieve, J, Biber, D, Friginal, E. and Nekrasova, T. 2010. *Variation Among Blogs: A Multi-dimensional Analysis*. In Mehler,A., Sharoff,S., and Santini,M. (eds.), *Genres on the Web: Computational Models and Empirical Studies*. Springer: New York.

Herring, S.C., Scheidt, L. A., Bonus, S., Wright, E. (2004). Bridging the Gap: A Genre Analysis of Weblogs. *Proceedings of the 37th Hawaii international conference on system sciences (HICSS-37).* Los Alamitos: IEEE Computer Society Press.

Herring, S. and Paolillo, J. (2006). Gender and Genre Variation in Weblogs. *Journal of Sociolinguistics* 10 (4), 439-459.

Herring, S. (2011). Computer-Mediated Conversation: Introduction and Overview. *Language@Internet* 8. Retrieved April 20, 2013, from http://www.languageatinternet.org/articles/2011/Herring

Herman, D., Jahn, M., and Ryan, M.L. (2005). *Routledge Encyclopedia of Narrative Theory*. Abingdon: Routledge.

Heylighen, F. (1993) Selection Criteria for the Evolution of Knowledge. *Proceedings of the 13th International congress on Cybernetics.*

Heylighen, F. and Dewaele, J. (1999). Formality of Language: definition, measurement and behavioral determinants. *Internal Report, Center "Leo Apostel",* Free University of Brussels.

Heylighen, F., and Dewaele, J. (2002). Variation in the Contextuality of Language: an Empirical Measure. *Foundations of Science*, 7 (3), 293–340.

Horowitz, R., and Samuels, J. S. (1987). Comprehending oral and written language: Critical contrasts for literacy and schooling. In R. Horowitz and J.S. Samuels (Eds.), *Comprehending oral and written language* 1-52. New York: Academic Press. In Herring, S. (2011). Computer-Mediated Conversation: Introduction and Overview. *Language@Internet* 8. Retrieved April 20, 2013, from http://www.languageatinternet.org/articles/2011/Herring

Hyland, K. (1996). Writing without conviction? Hedging in scientific research articles. *Applied Linguistics* 17(4): 433–454. In Knight, D., Adolphs,S., Carter, R. (2013). Formality in Digital Discourse: A study of Hedging in CANELC. *Yearbook of Corpus Lingusitics and Pragmatics 2013: New Domains and Methodologies.* Springer, 131-152.

Knight, D., Adolphs,S., Carter, R. (2013). Formality in Digital Discourse: A study of Hedging in CANELC. *Yearbook of Corpus Lingusitics and Pragmatics 2013: New Domains and Methodologies.* Springer, 131-152.

Lahiri, S., Prasenjit M., and Xiaofei, L. (2011). Informality Judgment at Sentence Level and Experiments with Formality score. In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing, 2*.

Lee, D. (2001). Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path Through the BNC Jungle. *Language Learning & Technology* 5 (3) 37-72. Retrieved April 25, 2013 from http://llt.msu.edu/vol5num3/lee/default.html

Leech, G. and Svartvik, J. (2002) *A Communicative Grammar of English*. Essex: Pearson Education Limited.

Levelt ,W.J.M. (1989) *Speaking. From intention to articulation*, Cambridge: MIT Press. In Heylighen, F. and Dewaele, J. (1999). *Formality of Language: definition, measurement and behavioral determinants.* Internal Report, Center "Leo Apostel", Free University of Brussels.

Levinson, S. (1983). *Pragmatics*. Cambridge: University press.

Ling, R. 2003. The socio-linguistic of SMS: An analysis of SMS use by random sample of Norwegians. In Mobile communications: Renegotiation of the social sphere, ed. R. Ling and P. Pedersen, 335–349. London: Springer. In Knight, D., Adolphs,S., Carter, R. (2013). Formality in Digital Discourse: A study of Hedging in CANELC. *Yearbook of Corpus Lingusitics and Pragmatics 2013: New Domains and Methodologies.* Springer, 131-152.

Miller, C.R. (1984). Genre as social action. *Quarterly Journal of Speech*, 70,151-167

Myers, G. (2010). *Discourse of Blogs and* Wikis. London: Continuum International Publishing Group.

Maynor, N. (1994). The language of electronic mail: Written speech? In G. Little & M. Montgomery (Eds.), *Centennial usage studies. American Dialect Society, 78,* 48-54. In

Herring, S. (2011). Computer-Mediated Conversation: Introduction and Overview. *Language@Internet* 8. Retrieved April 20, 2013, from http://www.languageatinternet.org/articles/2011/Herring

Nowson, S., Oberlander, J., and Gill, A.J. (2005) Weblogs, genres and individual differences. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society,*1666-1671.

Nowson, S. (2006). *The Language of Weblogs: A study of genre and individual differences.* Unpublished Doctoral Thesis. University of Edinburgh.

O'Reilly, T. (2007). What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software Communications & Strategies, *International Journal of Digital* Communication 65, 17. Retrieved May 14, 2013 from http://www.unige.ch/ses/socio/pdrs/programme/20072008/collectifsmorges/Communicationsstrategies.pdf

Quirk, R., Sidney, G., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language.* London: Longman

Rees, A. (1983). Pronouns of Person and Power: A study of Personal Pronouns in Public Discourse. Unpublished MA dissertation, Sheffield University. In Vaughan, E & Clancy, B. (2013). *Small Corpora and Pragmatics. Yearbook of corpus linguistics and Pragmatics 2013*. Springer, 53-73.

Sack, W. (2000). Conversation Map: An interface for very large-scale conversations. *Journal of Management Information Systems, 17*(3), 73-92. In Herring, S. (2011). Computer-Mediated Conversation: Introduction and Overview. *Language@Internet* 8. Retrieved April 20, 2013, from http://www.languageatinternet.org/articles/2011/Herring

Scott, M. (2004-2006). *Oxford Wordsmith Tools.* Version 5.0. Oxford: Oxford University Press.

Swales, J. M. (1990). *Genre Analysis English in Academic and Research settings*. Cambridge: Cambridge University Press.

Tagg, C. 2009. A corpus linguistics study of SMS text messaging. Unpublished Ph.D. thesis. Birmingham: The University of Birmingham. In Knight, D., Adolphs,S., Carter, R. (2013). Formality in Digital Discourse: A study of Hedging in CANELC. *Yearbook of Corpus Lingusitics and Pragmatics 2013: New Domains and Methodologies.* Springer, 131-152.

Teddiman, L. 2009. "Contextuality and beyond: Investigating an online diary corpus". In *Proceedings of the Third International Conference on Weblogs and Social media,* 331-333.

Toutanova, K., Klein, D., Manning, C., and Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT*-NAACL, 252-259.

Vaughan, E. and Clancy, B. (2013). *Small Corpora and Pragmatics. Yearbook of corpus linguistics and Pragmatics 2013*. Springer, 53-73.

Word press.com. *A live look at activity across wordpress.com.* Retrieved January 10, 2013 from http://en.wordpress.com/stats/

Yates, S. (1996). "Oral and Written aspects of computer conferencing" In Herring, S (ed),*Computer mediated communication: linguistics, social and cross-cultural perspectives.* Amsterdam:John Benjamins Publishing co.

Yule, G. 1996. *Pragmatics*. Oxford :Oxford University Press.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley.

Appendix A

F-scores of forum, blog, news and academic paper data set from Lahiri et al. (2011: 455)

Dataset	F-score
Forum	58.52
Blog	65.24
News	66.51
Academic Paper	68.62

Appendix B

Brown corpus word classes frequency per 1000 words

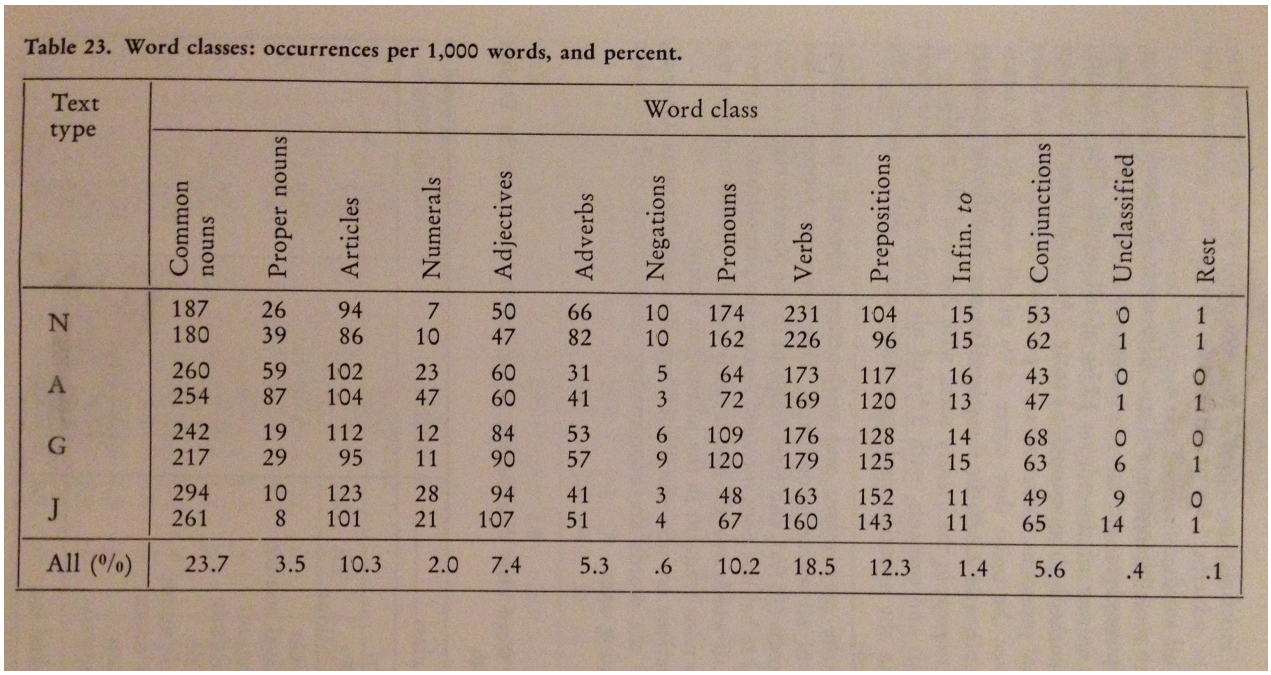

Table 23. Word classes: occurrences per 1,000 words, and percent.

Text type	Word class													
	Common nouns	Proper nouns	Articles	Numerals	Adjectives	Adverbs	Negations	Pronouns	Verbs	Prepositions	Infin. *to*	Conjunctions	Unclassified	Rest
N	187	26	94	7	50	66	10	174	231	104	15	53	0	1
	180	39	86	10	47	82	10	162	226	96	15	62	1	1
A	260	59	102	23	60	31	5	64	173	117	16	43	0	0
	254	87	104	47	60	41	3	72	169	120	13	47	1	1
G	242	19	112	12	84	53	6	109	176	128	14	68	0	0
	217	29	95	11	90	57	9	120	179	125	15	63	6	1
J	294	10	123	28	94	41	3	48	163	152	11	49	9	0
	261	8	101	21	107	51	4	67	160	143	11	65	14	1
All (%)	23.7	3.5	10.3	2.0	7.4	5.3	.6	10.2	18.5	12.3	1.4	5.6	.4	.1

Appendix C

Frequency of hedges in all the blog topics

Hedge	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
Just	2.9	4	3.8	2.1	3.45	1.55	2	2.7
Really	0.85	1.1	0.9	1.6	1.45	0.6	1	1.3
Only	0.75	1.65	0.75	1.3	1.1	2.35	1.2	1.95
Actually	0.4	0.8	0.4	0.25	0.05	0.35	0.5	0.15
Quite	0.15	0.4	0.35	0.2	0.25	0.45	0.1	0.2
Thing	0.7	0.35	0.55	0.2	0.5	0.55	0.6	0.95
Maybe	0.55	0.35	0.3	0.25	0.25	0	0.15	0.7
Probably	0.35	0.9	0.15	0.7	0.45	0.4	0.85	0.15
Of course	0.5	0.2	0.35	0.2	0.2	0.15	0.35	0.1
I think	0.25	1.15	0.85	0.45	0.6	0.25	0.35	0.7
You know	0.4	0.1	0.3	0.05	0.3	0.05	0.35	0.4
Likely	0.45	0.4	0	0.05	0.05	0.45	0	0.05
Apparently	0.15	0	0.05	0.05	0	0	0.35	0
Guess	0.1	0.2	0.25	0.15	0.1	0.05	0.55	0.05
Usually	0.05	0.15	0.2	0.05	0.35	0.1	0.05	0.15

Surely	0.25	0	0	0.05	0.05	0	0	0
Generally	0.05	0.25	0	0.05	0.1	0	0	0.1
Sort of	0.2	0.15	0.2	0	0.2	0.05	0.05	0.05
Normally	0	0	0.1	0.2	0.05	0.15	0.1	0.05
Kind of	0.35	0.2	0.25	0	0.45	0	0	0
Relatively	0	0.2	0	0.1	0.05	0.1	0.05	0.05
Possibility	0.05	0	0	0	0	0.05	0	0
Frequently	0.05	0	0.05	0	0.3	0	0.05	0
Necessarily	0.05	0.05	0	0.05	0.05	0.05	0.1	0
Roughly	0	0.05	0	0	0.05	0	0	0
Typically	0	0.05	0	0	0.05	0	0	0
Seemingly	0.05	0	0	0.05	0	0.05	0.05	0
Broadly	0	0	0	0	0	0.05	0	0
Arguably	0	0	0.05	0	0	0.1	0.05	0.05
Partially	0	0	0	0	0.05	0	0	0
Total	9.6	12.7	9.85	8.1	10.5	7.9	8.85	9.85

Appendix D

Frequency of private verbs in all the blog topics

Private verb	Politics	Finance	Art	Family	Food	Sport	Celebrity	Personal
Accept	0.05	0.05	0	0	0.05	0.05	0	0.05
Anticipate	0	0	0	0	0	0	0.05	0
Ascertain	0	0	0	0	0	0	0	0.05
Assume	0.1	0.05	0	0	0	0	0.15	0
Believe	0.75	0.1	0.15	0.3	0.05	0.3	0.4	0.25
Calculate	0.05	0	0		0	0	0	0
Check	0.15	0.55	1.05	0.6	0.2	0.2	1.4	0.1
Conclude	0	0	0	0.05	0	0.05	0.05	0
Conjecture	0	0	0		0	0	0	0
Consider	0.25	0.15	0.15	0.2	0.15	0.15	0.1	0.05
Decide	0.05	0.1	0.4	0.35	0.05	0.05	0.45	0.55
Deduce	0	0	0	0	0	0	0	0
Deem	0	0	0	0	0	0	0	0

Demonstrate	0.15	0	0	0.05	0	0.05	0.05	0.05
Determine	0.1	0.1	0	0.1	0	0.25	0.05	0.05
Discern	0		0	0	0	0	0	0
Discover	0	0.55	0.1	0.05	0	0.05	0.05	0.15
Doubt	0.05	0	0	0.05	0	0	0.15	0
Dream	0.05	0.15	0.1	0.2	0.4	0	0.55	0.4
Ensure	0.05	0.25	0	0.1	0	0.1	0	0.05
Establish	0.05	0	0	0.15	0	0.25	0	0
Estimate	0	0.15	0	0	0	0	0.05	0
Expect	0.15	0.15	0.05	0.15	0.1	0.7	0.35	0.1
Fancy	0	0	0.4	0	0.35	0	0	0
Fear	0.35	0	0.4	0.05	0.05	0.1	0.1	0.1
Feel	0.25	0.65	0.4	1.7	0.5	0.6	0.4	0.6
Find	0.55	1	0.8	1.5	0.65	0.55	0.3	0.65
Foresee	0	0		0	0	0	0	0
Forget	0	0.1	0.1	0.15	0.1	0.15	0.2	0.2
Gather	0	0		0	0	0	0	0.1
Guess	0.15	0.2	0.25	0.2	0.1	0.1	0.55	0.05
Hear	0.75	0.2	0.5	0.35	0.2	0.05	0.05	0.75
Hold	0.25	0.4	1.55	0.45	0.15	0.55	0.2	0.3
Hope	0.35	0.2	1.55	0.35	0.2	0.3	0.65	0.5
Imagine	0.25	0.05	0.1	0.25	0.15	0	0	0.25
Imply	0	0	0	0	0	0	0	0
Indicate	0	0.05	0.05	0.05	0	0	0	0.05
Infer	0	0	0	0	0	0	0	0
Insure	0	0	0	0	0	0	0	0
Judge	0.2	0.1	0	0.05	0.05	0.1	0.55	0.05
Know	2.4	1.65	1.8	1.6	1.65	0.85	1.3	2.5
Learn	0.1	0.6	0.3	1.2	0.3	0.05	0.2	0.35
Mean	0.25	1.15	0.4	0.9	0.5	0.35	0.45	0.55
Note	0.4	0.25	0.05	0.25	0	0.1	0.05	0.25
Observe	0.1	0	0	0	0	0	0	0
Perceive	0	0	0	0	0	0.05	0	0
Presume	0	0.05	0	0	0	0.05	0	0
Presuppose	0	0	0	0	0	0	0	0
Pretend	0.1	0	0	0.05	0	0.05	0.2	0.05
Prove	0.1	0	0	0	0.05	0.1	0.2	0.05
Realize	0	0.1	0.2	0.15	0	0.05	0.15	0.15
Reason	0.15	0	0	0.15	0	0.5	0.35	0.3
Recall	0.05	0	0	0.05	0	0.15	0	0

Reckon	0	0	0	0	0	0	0	0
Recognize	0	0.05	0.1	0.1	0.05	0	0	0
Reflect	0	0.05	0	0	0	0	0.05	0.1
Remember	0.1	0.3	0.05	0.7	0.15	0.15	0.3	0.2
Reveal	0.25	0	0.35	0.05	0	0.15	0.7	0.05
See	0.85	1.4	1.7	1.1	0.75	0.85	0.8	2.45
Sense	0.2		0.05	0.05	0	0.35	0.05	0.15
Show	1.2	0.65	0.6	0.05	0.05	0.5	0.55	0.6
Signify	0	0	0	0	0	0	0	0
Suppose	0	0.35	0.05	0.05	0	0	0.2	0.05
Suspect	0.6	0.05		0	0.15	0	0	0
Think	1.25	2.7	1.7	1.4	1.1	1.1	1.95	2.65
Understand	0.2	0.2	0	0.95	0	0	0	0.3
Total	13.4	14.85	15.45	16.25	8.25	10.1	14.35	16.2