Register Variation in Modern Written Icelandic

Jim Wood

University of New Hampshire

Abstract

In this paper, I compare two online written registers of modern Icelandic: online blogs and online newspapers. I review some of the more successful attempts at characterizing register linguistically and give an overview of Biber's multi-dimensional techniques, where register is viewed as varying along several dimensions. I review the most salient of his dimensions, Dim. 1, Informational vs. Interactive, and examine some of the features which characterize it in English. I then show that these features act in Icelandic as in English: the more informational features are more prevalent in online newspapers and the more interactive features are more prevalent in online blogs. I also look at the distribution of two Icelandic-unique syntactic structures, impersonal topicalization in main clauses and stylistic fronting in relative clauses, and show that their variation depends on register. Newspapers, as the informational register, prefer a more compact delivery of information, and in both structures prefer the variant which requires movement. Blogs, on the other hand, prefer the syntactic variants which avoid movement.

I discuss the significance of featural-distribution analysis in corpora by describing a few recent studies which have used distributions to shed light on various aspects of generative grammar. I finish by describing register variation as an acquired language variety which exploits a basic grammar in ways that speakers/users acquire by exposure. That is, features such as elaborated noun phrases and preference for finite 'be' become tacit knowledge at various levels of competence on the part of the language-user. I give a number of reasons why Icelandic might be a logical next choice for a multi-dimensional study: it is similar to and different from English in theoretically crucial ways which could shed light on universals or universal tendencies of register variation.

1 Defining Register Variation

There are several factors which can cause or influence intra-language variation. Regional variation, or dialect, is characterized by the fact that people talk like the people they talk to most, and, of course, people talk most to those who are closest to them. Sociolinguistic variation represents cultural and demographic variation. It is also due the fact that people talk most like the people they talk to, but it focuses less on region and more on age, gender, socio-economic status, and sub-cultures. Register variation focuses on how language varies according to context. People use language differently in different situations, and the study of register variation attempts to capture the linguistic differences influenced by the situational context. Every speaker of every language uses a variety of registers in his everyday life.

1.1 Previous Research on Register

Registers have been studied in a variety of ways. Some register studies focus on a single register, and attempt to characterize that register descriptively. Such studies will often describe lexical particulars, such as “the count” in baseball broadcasts. Lexical particulars such as this are sometimes considered to be register markers, so if a transcript contains the phrase “the count is three and one”, it is probably a transcript of a baseball broadcast (Ferguson 1983). Studies of individual registers also look at particular linguistic features such as pronunciation variables (post-vocalic r-dropping dependant on attention to speech (Labov 1972)), or syntactic variables (clause complexity and/or the

presence of a particular syntactic structure, such as in Mardh (1980)).[1] Atkinson and Biber (1994) provide a thorough review of previous empirical research on register.

According to Biber (1988, 1995), it was not possible until recently to perform comprehensive register studies. This is because a comprehensive register study would contain the most linguistic features across the best representations of the most registers. Linguists have long acknowledged that it is not one or two linguistic features that characterize a register, but a collection of many co-occurring features. A conversation has a high level of interactivity, and thus contains many first and second person personal pronouns, but it also contains many other features that co-occur with personal pronoun usage in the same relative frequency. Before automated computational processes were developed, it was not feasible to analyze the types of features that co-occurred, for various reasons, at a statistically significant level.

In his groundbreaking 1988 study, Biber analyzed 67 linguistic features across hundreds of thousands of lines of text in many registers. He found six 'factors' in which the frequency of a set of linguistic features co-occurred. He called these six factors dimensions, and, though they were discovered linguistically, described them functionally: (1) Involved vs. Informational, (2) Narrative vs. Non-Narrative, (3) Situation-Dependant vs. Elaborated, (4) Overtly Argumentative vs. Non-Overtly Argumentative (or Non-Argumentative), (5) Abstract vs. Non-Abstract, and (6) Online Informational vs. Edited or Non-Informational. They served, thus, as a series of continua which, combined, could describe any given register. A register could be highly informational, not very narrative, situation dependant, very abstract, etc. According the Biber's study, for example, Radio Broadcasts and Romance Fiction are on complete opposite ends of the continuum of

[1] Ferguson (1983) looked at syntactic variables as well.

Dimension 2, Narrative vs. Non-Narrative, but do not differ as much in regards to features that characterize Dimension 1, Involved vs. Informational.

His 1988 study was the basis of many register analyses in English thereafter, and further study was recommended that could include different linguistic features and different registers. Connor and Upton (2003), for example, compiled a corpus of Non-Profit Organization Direct Mail Letters and used Biber's framework to calculate this particular genre's scores on five of the six genres. Among other findings, they found that, surprisingly, the corpus contained very little overt expression of argumentation.

Since linguistic features vary from language to language, and language use within a culture varies as well, each language requires a comprehensive analysis such as Biber's for English. Studies were thus done in Korean, Somali, and Nukulaelae Tuvaluan. The results were compiled in his 1995 work, which was the first attempt at a cross-linguistic comparison of register variation. Davies is currently working on a comprehensive study of register variation in Spanish.

There is some evidence in Biber's 1995 study that register variation surfaces differently from language to language in ways partially dependent on the structural properties of the language. Korean, for example, has an entire language-specific dimension, Dimension 6 "Honorification," which, although certainly motivated culturally, exists linguistically due to some features of the Korean language which do not exist in many other languages: three out of five features characterizing this dimension are honorific expression, humble expression, and 'formal' sentence ending. There is no opportunity for such a dimension in which English (or many other languages) could vary, since there are no equivalent linguistic features.

Contrarily, there are several dimensions that seem to have representations in most, if not all, languages. Dimensions having to do with interactivity, informational elaboration, and narration seem to surface no matter what the language in question is. Further, many of the linguistic features that characterize these dimensions are shared among languages. Biber has even suggested that linguistic similarities between similar registers of different languages are more striking than similarities between different registers of the same language.

Icelandic is structurally similar to English in many important ways. It is an SVO language, with almost identical parameter settings (that is, SPEC, HEAD, COMP). Adjectives precede the nouns they qualify, it has a separate infinitive marker *að* equivalent to English *to*, particle verbs are often very similar (such as *að telja upp* 'to count up'), and prepositions are seperate words that take noun phrase complements. Assuming that many of these features have the same function in Icelandic (prepostions and nouns, for example, give the same kind of information) as in English, it may be possible to study many of the characteristics of variation in Icelandic under the prediction that some of the linguistic features which seem to hold similar functions across very different languages will hold these same functions in Icelandic.

1.2 Situating Icelandic in the Register Landscape

Register variation, to date, has not been comprehensively studied in Icelandic. One study (Barðal 2003) did look at frequency distribution of a feature in Icelandic while considering register. In this study, morphological case was examined in regards to its relationship with syntactic and thematic contexts. Register, however, was not a primary

aim of the study. Rather, Barðal accounted for register while studying the overall language, and only made the distinction between written/spoken and formality, as follows: “The frequency of dative subjects, and also Experiencer subjects, shows variation according to registers, in that dative subjects are more common in spoken Icelandic than in written Icelandic, and almost non-existent in the more formal styles of written language” (pp. 182). The implications in terms of these conclusions were: “Thus, conclusions on oblique subjects in historical investigations have to take into account the type of genre being investigated” (pp. 183).

A multi-dimensional study of the kind done by, for example, Connor and Upton (on Direct Mail Letters) in 2003, requires a previous comprehensive multidimensional register study on which to base it. Unfortunately, such a study has not been done in Icelandic and the computational tools required for such a study are not yet available. The goal of the present study, thus, is to analyze and compare variation in two somewhat similar registers in Icelandic, under the hypothesis that many of the characterizations of linguistic features described for English (and the other languages) hold for Icelandic. For example, there is no reason to believe that personal pronouns are less interactive in Icelandic than they are in English, or, conversely, that long, elaborated noun phrases are more informational in English than they are in Icelandic.

A second goal of the present study is to look at the distribution of some syntactic structures in Icelandic that have no equivalent in English. I use the characterizations based on the distribution of linguistic features known to have certain functions to discuss the significance of the distribution of these syntactic structures. I distinguish between assuming functional purposes to distributions in the communicative sense and the

functionalism assumed in functionalist grammars. Studying the distribution of linguistic features can be used to shed light on the pragmatics of situationally-motivated language varieties. I also argue that analyzing meaningful distributions of linguistic features (nouns, for example) is not incompatible with a generative perspective; rather, such distributions are evidence for tacit register competency. In the next section, I will distinguish functional motivations of features from functional grammar perspectives.

1.3 Functional vs. Generative Grammar

Because an important aspect of quantitative studies on register variation such as this one is the functional motivations of linguistic features, it is important that I clarify what I mean by 'functional' and distinguish it from the theoretical frameworks known as functional grammars. Functional grammars investigate language from a particular perspective. There are some differences among the theories, but the most important points that distinguish them from generative grammar are their top-down approach and their rejection of innateness as formulated by generative grammarians. The top-down approach views the speech act as the primary unit of investigation, and therefore considers syntax, phonology, etc. to all be secondary. It sees abstract grammar as evolving from functional requirements (such as communication). This is in stark contrast to the generative approach which, of course, attempts to uncover the language production tool from the bottom up; that is, the rules that produce infinite language. The generative approach does not deny that there are pragmatic functions associated with language use (what Chomsky calls 'e-language'), but it sees those functions as stemming from the use of a biologically bestowed language organ. The functionalists see it the other way around.

Their idea of innateness consists, at most, of a social propensity towards language and communication on the part of the human animal.[2]

The functional properties of linguistic features, however, are not at odds with the generative perspective. Furthermore, the analysis of frequencies of features in corpora has in recent years been used by generative linguists to shed light on various aspects of internal grammar. One good example of a frequency study having direct grammatical implications is Barðal's (2003) study of morphological case. She presents a frequency count on a personally assembled corpus of 40,000 words of the morphological case of nouns, pronouns, and adjectives, relating them to thematic roles and syntactic functions. Her empirical research shows that both traditional and generative grammar fall short of adequately accounting for morphological case. That is, current notions of morphological cases' relationship to either syntactic functions or thematic roles require too many exceptions when examined in real use. This has real consequences in any grammatical description or theory, particularly in the area of language acquisition. If, for example, 25% of all direct objects in a language (in this case, Icelandic) are in the dative case, as is shown in her study, both syntactic and lexical accounts of grammar wind up rather inelegant. Note that without a frequency of this kind, it is much easier to account for a majority construction (accusative case objects) and deal with the minority construction (dative case objects) in an exceptional or "problematic" way.

Cornips and Corrigan (2005) compiled a collection of papers intended to allow variationists and generativists to meet on common ground. Adger and Smith, for example, argue that the Minimalist Program can be efficiently employed to account for

[2] Bybee (1998) gives an account of how the evolution of abstract grammar is accounted for from the functionalist perspective. Edmondson and Burquest (1998) give a basic overview of functionalist theory.

variation. They use corpora to look at variation in several syntactic variables, such as number agreement, distinguishing between agreement which is categorical (invariable) and variable. They argue that within the Minimalist Program account of grammar, variable agreement can be accounted for by whether or not a 'tense head' carries an uninterpretable number feature. Speakers have a choice between selecting T1, which has an uninterpretable number feature and T2, which does not. The choice, which becomes as simple as a lexical choice, determines whether or not there will be agreement. Such a study makes predictions about both grammar (in the generative sense) and variation, but could not be conducted without empirical research using distributions (in this case, of when agreement is variable/invariable) to distinguish between categorical and variable levels of grammar.

Another study which makes use of corpora and Minimalism is van Gelderen (2005). She uses multiple corpora and frequencies of conjunctions (distinguishing between Pronoun and Noun) to hypothesize about the internal structure of noun phrases and determiner phrases. Namely she argues that, at least since 1600, first person personal pronouns are more likely to be heads of a phrase than nouns are. She also uses corpus frequency to hypothesize that accusative case is the default case, and explains sentences such as *Me and him went to the store* in terms of structural economy (using default case in coordinates rather than checking case in structure). Again, she uses frequencies to come to conclusions which are not by any means at odds with generative grammar.

The goal of a register study is to characterize a linguistic situation functionally or pragmatically. It is thus useful, when trying to characterize the linguistic effect of a situation, to examine the frequency distributions of linguistic features which could,

potentially, have a meaningful effect on how that situation is linguistically perceived. Taking the functional use of (for example) prepositions and examining their frequencies in context, thus, does not have to be situated within the theoretical framework of functional grammars, but can follow directly from a generative perspective. Further, studies such as Barðal (2003), Adger and Smith (2005), and van Gelderen (2005) show how the distributions of such features can contribute to discussions of grammar within the generative framework.

1.4 Relevant Syntactic Structures

In this section I will give an overview of the aspects of Icelandic syntax discussed in this paper. I will be focusing mostly on Stylistic Fronting (SF) in relative clauses and Topicalization vs. Expletive Insertion in impersonal main clauses, although I will touch on a few other word order issues as necessary.

The basic sentence structure is Subject-Verb-Object, but there is some variation in where constituents can end up. Icelandic is a fairly strict verb-second (V2) language, although V1 occurs as well in a phenomenon known as Narrative Inversion. Examples of V2 are clear in cases of Topicalization, where a non-subject constituent is fronted to the beginning of the clause. The verb stays in place, and the subject, if there is one, follows the verb. Examples of this are in (1) below.[3] As we can see, the fronted constituent can be an adverbial (a), an object (b), a prepositional phrase (c), or a negative (d).

1. a. Núna hef ég afsannað það.
   Now have I disproved that.
   Now I've disproved that.

[3] All examples on word order in this section unless otherwise noted are taken directly from the blog and newspaper corpora.

b. Eitthvað hef ég heyrt um það
Something have I heard about that
I've heard something about that.

c. Fyrir miðjan júlí verði innan við 500 manns eftir í stöðinni.
Before mid-July will-be in with 500 people after in the-base
By mid-July, there will be fewer than 500 people left on the base.

d. Ekki er lengur þörf á vopnaðri uppreisn Íraka.
Not is anymore need on armed insurgence of-Iraq.
There is not a need for armed insurgence in Iraq anymore.

Verbal particles and participles are also sometimes possible. Note that only adverbials, PPs, and objects can be fronted in cases where an overt subject is present. In cases such as (1d), or in cases of fronted verbal particles and participles, it must be an impersonal clause. This is because fronted constituents of this kind require a subject gap. The gap can be filled by either a fronted constituent (of any kind) or by the insertion of an expletive. Main clauses must fill this subject gap. The choice between fronting and expletive insertion is shown below in (2):

2. a. Það var verið að spyrja um hvort...
(exp.) was been to ask about whether
It had been asked whether...

b. Verið er að yfirfara allt kjarnorkuverið.
Been is to inspect all the-nuclear-power-station
All the nuclear power station has been inspected.

Notice the position of *verið* 'been' in each sentence. In (2a), it is post-verbal, in its underlying position. In (2b), it is preverbal, fronted to fill the subject gap. It would be equally grammatical to swap these structures in each case, which would result in (3a) and (3b):

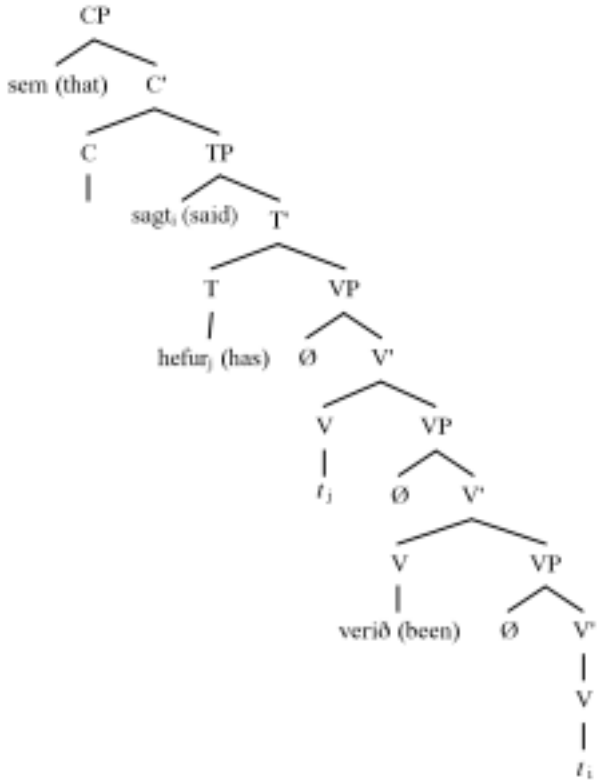
3. a. Verið var að spyrja um hvort...

b. Það er verið að yfirfara allt kjarnorkuverið.

The choice between these structures will occupy the bulk of our discussion of syntactic distributions in sections 3.2 and 4.2.

Another important choice in Icelandic syntax is that of Stylistic Fronting in embedded clauses. Here, I will be taking the analysis of Maling (1990) and (mostly) Hrafnbjargarson (2004) that SF in relative clauses is the result of filling the subject gap in the clause. I assume that, when fronted, the constituent occupies SPEC,TP as shown in Figure 2.1 below.[4]

Figure 1.1 Stylistic Fronting in Relative Clauses[5]



[4] A alternative analysis (that of adjunction) is discussed in Holmberg (2000), where he points out some of its weaknesses. Hrafnbjargarson (2004) also point problems with this alternative analysis, and it is his and Maling´s (1990) analysis that we will assume in this paper.

[5] I left the SPEC,VP slots open because they may be instrumental in allowing the participle to move up the tree to SPEC,TP, regarding the Minimal Link Condition.

The speaker/writer has a choice between performing the movement operation as illustrated above, resulting in (4a) below, and leaving the participle in its underlying position, indicated as t_j in figure 2.1 above, resulting in the surface form shown in (4b) below.

4. a. ...sem sagt hefur verið
that said has been
...that has been said

b ...sem hefur verið sagt
that has been said
that has been said

The choice between (1) SF and no-SF in relative clauses and (2) Topicalization and Expletive Insertion in main clauses will be the main points discussed in sections 3.3 and 4.3.

1.5 Description of the Registers

Key to Biber's (1988, 1995) findings in multi-dimensional register studies is the idea that there is no direct dichotomy between speech and writing; a prepared speech may be in many ways more like an academic article than a conversation, even though the former is written and the latter spoken. Medium, therefore, does not determine the register directly. Preparation time and opportunity for editing, however, can have a dramatic effect on the linguistic character of a register. Still, written and spoken language often have very different contexts in which they are used, and the extra-linguistic variables are also markedly different. Crystal (2001) hypothesizes that language on the internet is influenced by extra-linguistic contexts so markedly different from either traditional writing or speech, that it should be considered its own "production medium":

"Whereas in the past we have had speech, then writing, and throughout the 20th century debated the relationship between the two, now we are faced with a new medium, and one which could be bigger than either of its predecessors" (pp. 241). It, too, has its own spectrum of registers ranging from e-mails to websites, each with its own set of extra-linguistic influences which affect the production setting and therefore shape and characterize the registers.

The registers I am considering in the present study are both internet registers, and they are both in Icelandic: online newspapers and online blogs. They are similar in several ways. First, they are both characterized by a large collection of smaller entries, which are meant to tell something to a wider audience, although the entries in blogs vary in size more than those in newspapers. Second, they are produced by roughly the same age group, again with blogs being more variable. Hrafnbjargarson, an Icelandic linguist, has pointed out to me in a personal communication (Jun 1, 2006) that online newspapers are not direct representations of their printed counterparts, but rather are written by young college students, usually aged 18-22. They also, according to Hrafnbjargarson, often use non-standard language. This is similar to blogs in the sense that they are written in potentially non-standard language by young people.

They also differ in some important ways. First, although online newspapers are written in potentially non-standard language by young speakers, the impetus for editing is much larger than with blogs. Blogs tend to be written "on the go" with little to no proofreading or editing. Secondly, online newspapers are still imitating traditional newspapers and are subject to the same rules against libel, resulting in "source" phrases (such as *samkvæmt* 'according to') that are not necessary in blogs. Thirdly, they share a

purpose with traditional newspapers, namely, reporting. Within reporting, it is considered necessary to stay as objective as possible, so there is a sense that newspapers, although online and subject to e-mailed "letters to the editor" and, perhaps, e-mails to bylined reporters, are far less interactive than online blogs. Contrarily, bloggers expect to be contacted by their readers. In fact, bloggers expect that most people who read their blogs probably know them in person. They don't have to be objective at all; rather, making statements of pure opinion is encouraged and expected.

2 Methodology

In this section I describe the methodology used for this study. In section 2.1 I describe how I assembled the corpora. I will then discuss my methodology for collecting Dimension 1 Informational/Interactive features, analyzing lexical variation, and looking at distributions of syntactic structures in sections 2.2, 2.3, and 2.4.

2.1 Methodology for Assembling the Corpora

I assembled two corpora for this study. The first is a collection of blogs, the second, of online newspaper articles. I assembled my blog corpus by using search engines to search strings such as *er búinn að (gera e-ð)* 'lit. am finished to', a conversational way of saying 'have (done sth)'. This returned a high number of blogs. I then picked seven such blogs, written by four females and three males, to avoid gender bias. I took an equal amount of text from each, totaling 9,444 words. I used Microsoft Word's word count feature, after manually removing any characters that could result in an inaccurate count. I also removed blog entry markers, such as dates and *posted by*…,

along with anything which had no potential of being counted linguistically throughout the study. I allowed the words to be counted as the writer intended them, following the lead of Barðal (2003), so that when two words were written as one (as in *þarsem* for *þar sem* 'where'), they were counted as one. Abbreviations such as *t.d., til dæmis,* 'for example' were also counted as one word. The word counts for each blogger are shown in Table 2.1.

Table 2.1

Blog	Word Count
Elías	1423
Larusson	1281
Arnor	1425
Kolbeinn	1319
Sigrun	1219
Alma	1391
Stína	1386
Total	9444

The newspaper corpus was assembled somewhat differently, over a period of several weeks. I chose three online newspaper websites: Morganblaðið www.mbl.is, Visir www.visir.is, and Suðurland www.sudurland.is. The first two are very large and general, and the third is of a more local variety, related specifically to southern Iceland. I gathered 4,015 words of text from Morganblaðið, 4,219 from Visir, and 1,506 from Suðurland. Morgunblaðið and Visir were each divided into three sections. The number of words for each section are shown in Table 2.2. The total number of words in this corpus is 9,740. These word counts are also after removing all characters that could not be counted in linguistic frequency tests (such as dates and dividers) but would affect the word count function.

Table 2.2

Newspaper Word Count	
Morganblaðið Foreign	1823
Morganblaðið Domestic	1758
Morganblaðið Science/Technology	434
Visir Foreign	1848
Visir Domestic	1814
Visir Science/Technology	557
Suðurland	1506
Total	9740

2.2 Methodology for Dimension 1 Features

Once the corpora were assembled, I chose some linguistic features to count. The underlying assumption in Biber's feature-based register studies is that co-occurring linguistic features have functional motivations.[6] So the first step was to examine the features that characterize a certain functional continuum (dimension) and predict the level to which these features would vary in the Icelandic registers under consideration. The most marked dimension in English is the first, Involved vs. Informational, in the sense that this continuum is characterized by more co-occurring features than any other dimension. In addition, this register has the highest cross-linguistic universality and seems intuitively likely to be the biggest difference between blogs and newspapers. Both of these registers have a certain level of informational function, but we can predict that the online newspapers will be closer to the informational end of the spectrum, whereas blogs will be more interactive.

Biber's register studies characterize linguistic features by statistically positive and negative presence. A text with more positive features will be likely to have fewer of the negative features; they are in complementary distribution. Along Dimension 1, a text with

[6] At least in the sense discussed in section 1.3.

more positive features than negative is likely to be more interactive, whereas one with more negative features is likely to be more informational. Thus, we can expect that online newspapers will have a higher presence of the negative features and blogs will have a higher presence of the positive features. The terms 'positive' and 'negative' refer to statistical co-occurrence, though, and will not be considered as such in this study since I was not able to study statistical co-occurrence in the multi-dimensional sense of Biber (1988, 1995). Instead, I will refer to features as interactive (for positive) and informational (for negative).

I chose six interactive features to look for. I describe the particulars of how I looked for each feature and weighed them in section 3.1. The interactive features I looked for were: *vera* 'be' as finite verb, analytic negation *ekki* 'not', hv-questions (as in English wh-questions), demonstrative pronouns, indefinite pronouns, and first- and second-person personal pronouns. I then chose six informational features: regular nouns (as opposed to pronouns), word length, prepositions, type-token ratio, attributive adjectives, and location adverbials.

I automated the search process as much as possible, but, as described blow (section 3.1), there were some features that simply could not be searched comprehensively. For these features, I compiled two smaller (~1000 word) corpora in which to search manually and exhaustively. But in order to count NPs of a particular type (e.g. indefinite, demonstrative, and personal pronouns), it is useful to be able to estimate the total number of NPs for each corpus, in order to have a baseline with which to compare any given results. For example, let's say I am looking at indefinite pronouns. I find 73 in one corpus and 153 in another. In this case, I am including indefinite modifiers

and pronouns, so *some pigeon* would count as an indefinite noun phrase as much as *something* would. The total number of occurrences would best be weighed against the logical possibility of occurrence, rather than a more arbitrary number such as total corpus size in words. Thus, the logical possibility is that every noun phrase is indefinite; all sentences would be of a nature such as *some man told someone something about nobody.* In this case, we have a seven word sentence with four noun phrases, giving a 4:7 ratio of indefinite NPs to words. Since 4:7 would be the highest score in this case, it would be more useful to count this sentence as 1:1 ratio, indefinite NPs:NPs.

Of course, in order to come up with a more accurately estimated ratio, we would need more than an example sentence that exists in a vacuum. It is therefore more useful to see what ratio of NPs to words we find in the corpora themselves, and conduct some tests to see whether or not we can expect the ratio we find to continue throughout the corpora. To this end, I isolated five sections of the blogs and three sections of the newspapers and counted the NPs in each. If this were a sample size large enough to expect it to continue, we would find no significant difference between each section. Table 2.3 below shows these five sections, with the number of NPs per a certain number of words. The ratio is calculated in the next column. The totals are also calculated and the average ratio is presented at the bottom. Table 2.4 uses the average ratio (1 NP per 3.54 words) to predict the number of NPs that should appear in each section. The next column presents the difference between the prediction and the real number. Although there is some difference in these predictions, the average difference, as shown at the bottom of Table 2.4, is only 2.06. That is, the predicted number of NPs in these sections of the blog

corpus is only off from the actual number by two; it would thus be safe to estimate the total number of NPs in the blog corpus by dividing the total number of words by 3.54.

Table 2.3 Blog Corpus

	Actual NPs	Per # Words	Ratio 1:X
Section 1	137	448	3.27
Section 2	184	611	3.32
Section 3	107	383	3.58
Section 4	109	450	4.13
Section 5	122	419	3.43
Total	659	2311	3.51
Average			3.54

Table 2.4 Blog Corpus

	Actual NPs	Predicted	Difference From Actual
Section 1	137	126.55	10.45
Section 2	184	172.60	11.40
Section 3	107	108.19	-1.19
Section 4	109	127.12	-18.12
Section 5	122	118.36	3.64
Total	659	652.83	6.17
Average Difference			2.06

Tables 2.5 and 2.6, respectively, show the same for the newspaper corpus. The newspaper corpus, however, was a lot more stable, and only three sections were checked. The result was a slightly tighter ratio (1 NP per 3.07 words) and an average predictive difference of only -1.09.

Table 2.5 Newspaper Corpus

	Actual NPs	Per # Words	Ratio 1:X
Section 1	172	525	3.05
Section 2	103	298	2.89
Section 3	157	508	3.24
Total	432	1331	3.08
Average			3.07

Table 2.6 Newspaper Corpus

	Actual NPs	Predicted	Difference From Actual
Section 1	172	171.26	0.74
Section 2	103	97.21	5.79
Section 3	157	165.71	-8.71
Total	432	434.18	-2.18
Average			-1.09

Thus, we will estimate that there are 2668 NPs (9444/3.54) in the blog corpus and 3183 NPs (9740/3.07) in the newspaper corpus. We will use these figures as a benchmark for considering other types of NPs.

To return to the indefinite pronoun question, we will assume that any given noun phrase is either indefinite or other. I used Simple Concordance Program[7] to search both of the corpora for all morphological forms of a finite list of indefinite NPs.[8] I entered the various inflectional forms manually. The results are summarized in Table 2.7 below.

Table 2.7 Indefinite Pronoun Distribution

	Indefinite Pronouns	Other NPs	Total
Blogs	153	2515	2668
Newspapers	73	3110	3183

This distribution is significant: (Chi-square = 46.29 $p < .001$ d.f. =1).

These data have been presented here only as an example of how I use my NP estimations. They are presented again in the next section along with the other data.

[7] Version 4.09, from http://www.textworld.com/

[8] The indefinite nouns I used were as follows: *annar* 'another', *fáeinir* 'several', *enginn* 'nobody', *neinn* 'anyone', *ýmis* 'several', *báðir* 'both', *sumur* 'some', *nokkur* 'someone', *einhver* 'someone', *sérhver* 'everyone', and *hvorugur* 'neither'.

2.3 Methodology for Lexical Variation

To characterize lexical variation, I used Simple Concordance Program to do a series of frequency counts. This was a fairly simple and mostly automated process. I first took each corpus individually and performed a type-token ratio search. I then looked at the ten highest frequency words for each and compared the ones that were missing from each respective corpus. The data are presented in section 3.2 below.

2.4 Methodology for Syntactic Variation

In order to investigate the word order variation of the type described in sections 3.3 and 4.3, I first isolated all of the finite verbs in my corpora. I divided them into three categories: (1) Subject Precedes Verb, (2) Subject Follows Verb, and (3) No Overt Subject. I removed all relative and comparative clauses for separate consideration. The results for each corpus were very close numbers in each corpus, as shown in Table 2.8 below.

Table 2.8

	Blogs	Newspapers
Subject-Verb	57.5%	63.7%
Verb-Subject	22.6%	17.7%
No Subject	19.9%	18.6%

I then eliminated all instances that were of no interest. These included hv- and yes/no-questions (which would result in VS), finite verbs that were part of a conjoined VP (which would result in a particular structure being counted twice), and all types of imperatives (which would result in either VS or no overt subject). The modified results are shown in Table 2.9.

Table 2.9

	Blogs	Newspapers
Subject-Verb	61.8%	66.3%
Verb-Subject	23.5%	18.5%
No Subject	14.7%	15.2%

They have, as Table 2.9 shows, very similar distributions. A striking difference was found, however, when I examined the structures of the subjectless clauses.

Table 2.10 Types of Subjectless Clauses

	Blogs	Newspapers
Subject Dropping	47.8%	15.4%
Expletive	33.1%	5.9%
Impersonal	19.1%	77.9%
Long Distance Extracted Subject	0.0%	0.7%

As Table 2.3 shows, the majority of subjectless clauses in the blogs were due to what I call Subject Dropping. This is as ungrammatical in Standard Icelandic as it is in English. It happens when the subject of a clause is so clear that it is omitted. It only occurs in highly cohesive discourse and can be thought of as equivalent to English expressions as in (5) below, which is only grammatical when the subject is so obvious that it is omitted.

5. Went to the store yesterday.

The fact that it shows up so often in Icelandic blogs is probably due to (1) subjects often agree with person (though not as reliably as in, for example, the Romance languages) and (2) the highly ego-centric nature of blogs. That is, as shown by the high number of first-person personal pronouns (discussed in section 3.2 below), it is obvious enough when a first-person pronoun is omitted that the intended subject is *ég* 'I'. Similarly, the only time it is omitted in newspapers is either when the expletive *það* is not generated and is generally understood or in headlines where the subject is clear (or soon to be clear in the

article). The high amount of subject dropping in blogs is not especially interesting outside of the fact that it would add even more to an already high number of clauses where the first person is the subject. It is therefore of interest in characterizing blogs, as in section 3, but not of much syntactic interest in this section. We could either consider such clauses to be incomplete or as having some kind of *pro* as its subject. I will not consider subject dropping any further.

The other clause I want to consider and dismiss is one which I describe as long distance subject extraction. Here, the subject was doubly extracted from the clause; that is, there was one instance of a clause where the subject was extracted from a subordinate clause embedded within a relative clause. This sentence is given in (6) and will not be considered further here. Thus, the subordinate clause shown in italics below is eliminated from consideration in subjectless clauses.

6. …einkum þeim sem hann sagði *að hefði verið „slátrað“ fyrir baráttu þeirra fyrir mannréttindum.*

...also those who he said *that had been slaughtered for fight their for human-rights*

‘Also those whom he said had been slaughtered for fighting for human rights.’

3 Data

This section will present the data as well as some additional information regarding how the data were collected. Section 3.1 describes the Dimension 1 features, 3.2 describes Lexical Variation, and section 3.3 describes the distribution of the syntactic features discussed above.

3.1 Data for Dimension 1 Features

This section describes how I looked for and weighed the Dimension 1 features characterized as either informational or interactive. I present these data here and give any relevant information regarding their collection. It is divided into sections 3.1.1 and 3.1.2, where I present the distributions for (respectively) interactive and informational linguistic features.

3.1.1 Interactive Features

The interactive features I looked for were: *vera* 'be' as a finite verb, analytic negation *ekki* 'not', hv-questions (as in English wh-questions), demonstrative pronouns, indefinite pronouns, and first- and second-person personal pronouns. I looked for *vera* 'be' using Simple Concordance Program and entering all inflectional forms: all persons, past and present, indicative and subjunctive. I weighed the counts against the total number of finite verbs, which I counted manually, as described in section 2.3 above. The results are presented in Table 3.1 below.

Table 3.1 Finite *Vera* 'Be'

	Finite 'Be'	Other Finite Verbs	Total
Blogs	472	729	1201
Newspapers	313	803	1116

This distribution is significant: (Chi-square = 32.70 $p < .001$ d.f. =1).

I used the total finite verb count as a yardstick in two other features, analytic negation and hv-questions. Analytic negation is the way of negating the finite verb as in (1), and as opposed to (2) below. The difference in Icelandic is much the same as in English. In (1), the negative word *ekki* is a clause level negative which goes with the finite verb. In (2),

the negative word *enga* is an indefinite modifier (adjective) which declines with and modifies the noun (in this case *spurningu* ‘question’).

1. Ég hef ekki spurningu.
   I have not question
   I don´t have a question.

2. Ég hef enga spurningu.
   I have no question.
   I have no question.

To search for instances of this, I only needed to search for instances of *ekki* (which is indeclinable), a fairly easy process in Simple Concordance Program. The results are presented in Table 3.2 below. Here, ‘Other Finite Verbs’ is the number of verbs not negated in this way.

Table 3.2 Analytic Negation *Ekki* ‘Not’

	Analytic Negation	Other Finite Verbs	Total
Blogs	119	1082	1201
Newspapers	56	1060	1116

This distribution is significant: (Chi-square = 19.81 $p < .001$ d.f. =1).

The last interactive feature which I weighed against the number of finite verbs was hv-questions. These are almost identical to wh-questions in English, and they all begin with *hv* except for *af hverju* ‘why’: *hver* ‘who’, *hvað* ‘what’, *hvenær* ‘when’, *hvar* ‘where’, *hvers vegna* ‘why’ and *af hverju* ‘why’. These are sometimes declinable, so I searched all morphological forms in Simple Concordance Program and manually eliminated those that were not questions. The results were on a much smaller scale than the first two, and are presented in Table 3.3. Here, ‘Other Finite Verbs’ could be described as non-hv-question verbs.

Table 3.3 HV-Questions

	HV-Questions	Other Finite Verbs	Total
Blogs	20	1181	1201
Newspapers	0	1116	1116

This distribution is significant: (Chi-square = 18.75 $p < .001$ d.f. =1).

The other three interactive features were all weighed against the total NP count described in the methodology section above. The first of these was demonstrative pronouns. To search for demonstrative pronouns in all morphological forms, I used the prefix function and searched *þenn, þett* and *þess,* which account for the beginnings of all forms. I then went through the results by hand and eliminated all non-tokens and instances of demonstrative determiners (which have the same form, but modify a noun). Thus I was left with a count of all instances where an NP was made up of only these pronouns, and I weighed them against all other NPs, as in Table 3.4 below.

Table 3.4 Demonstrative Pronouns

	Demonstrative Pronouns	Other NPs	Total
Blogs	70	2598	2668
Newspapers	22	3161	3183

This distribution is significant: (Chi-square = 35.02 $p < .001$ d.f. =1).

The next NP type was indefinite NPs, which I already described in some detail above. I performed an extensive search of a finite list of indefinite nouns, and the results are repeated here for convenience's sake in Table 3.5.

Table 3.5 Indefinite Pronouns

	Indefinite Pronouns	Other NPs	Total
Blogs	153	2515	2668
Newspapers	73	3110	3183

This distribution is significant: (Chi-square = 46.29 $p < .001$ d.f. =1).

Lastly, I searched for personal pronouns. I again looked at all morphological forms of all first- and second-person personal pronouns, and I divided the results into each category and weighed them against 'other NPs,' as in Table 3.6 below.

Table 3.6 Personal Pronouns

	1st Sg	1st Pl	2nd. Sg	2nd. Pl	Other NPs	Total
Blogs	456	79	9	25	2099	2668
Newspapers	7	20	0	0	3156	3183

This distribution is significant: (Chi-square = 563.41 $p < .001$ d.f. =4).

3.1.2 Informational Features

I chose six informational features: regular nouns (as opposed to pronouns), word length, prepositions, type-token ratio, attributive adjectives, and location adverbials.

For regular nouns, I used the mini-corpora and counted manually. Table 3.7 below shows the number of regular nouns weighed against the number of 'other words'. These 'other words' include other parts of speech along with pronouns.

Table 3.7 Nouns

	Nouns	Other Words	Total
Blogs	204	906	1110
Newspapers	350	718	1068

This distribution is significant: (Chi-square = 59.45 $p < .001$ d.f. =1).

For word length, I used Simple Concordance Program's automatic calculation for the mean word length of each corpus. The results are presented in Table 3.8.

Table 3.8 Mean Word Length

	Characters	**Words**	**Mean Word Length**
Blogs	44600	9444	4.72
Newspapers	54462	9740	5.59

Prepositions were weighed against NPs. I used the 1000 word samples,[9] collected as described above, and counted prepositions manually. Since most prepositions take NP complements, I thought it made sense to weigh them against NPs, resulting in number of prepositions vs. non-prepositional NPs. The results are presented in Table 3.9 below.

Table 3.9 Prepositions

	Prepositions	Non-Prepositional NPs	Total
Blogs	118	196	314
Newspapers	157	191	348

This distribution is significant: (Chi-square = 3.86 $p < .05$ d.f. =1).

Type-token ratio was analyzed automatically by Simple Concordance Program. The raw numbers are indicated in Table 3.10, where the number of lexical items can be compared to the total number of words per corpus. I also combined both corpora into one and analyzed it in order to compare shared vocabulary, a subject which I return to in section 4.

Table 3.10 Type-Token Ratio

	Lexical Entries	Total Words
Blogs	2726	9444
Newspapers	3612	9740

Attributive adjectives were also counted in the smaller corpora, and then weighed against the NPs there. The results are presented in Table 3.11.

Table 3.11 Attributive Adjectives

	Attributive Adjectives	Other NPs	Total
Blogs	27	287	314
Newspapers	63	285	348

This distribution is significant: (Chi-square = 12.69 $p < .001$ d.f. =1).

[9] Blog Corpus: 1110; Newspaper Corpus: 1068

For location adverbials, I took a finite list based on the one used by Biber (1988 pp. 224) for English, and I searched for appropriate Icelandic equivalents.[10] I also considered the words that would have morphological variation, and accounted for that as well. I then went through the results by hand and removed any non-tokens. I weighed the results against finite verbs, an admittedly weak measurement, but the best I could come up with. Assuming far fewer adverbials than finite verbs, I assigned a 1:1 ratio of adverbials to verbs, subtracting the number of adverbs from the number of finite verbs. Even though some clauses, in reality, may have two or more adverbials along with one finite verb, this can still be an appropriate measurement if it is done the same way in both corpora, allowing us to compare them conveniently. Table 3.12 below shows the results of this.

Table 3.12 Location Adverbials

	Location Adverbials	Other Finite Verbs	Total
Blogs	47	1154	1201
Newspapers	87	1029	1116

This distribution is significant: (Chi-square = 16.00 $p < .001$ d.f. =1).

[10] The finite list is as follows: *um borð; fyrir borð* 'on board', *að ofan; uppi; uppi yfir; fyrir ofan; að framan; ofan við* 'over', *þvert yfir; að þvermáli* 'across', *breiður á undan; framundan; fram fyrir; fram á við* 'ahead', *við hliðina á; upp að* 'next to', *í kring; umhverfis; til meðvitundar* 'around', *að landi; í land; í landi* 'ashore', *aftur í skut; aftur á bak; fyrir aftan burt* 'behind', *í burtu* 'away', *að baki; á eftir; á eftir áætlun; á bak við; á eftir* 'behind', *niðri; að neðan; neðan við; fyrir neðan; niður fyrir; undir; niður á við; niður; niður stiga; niðri; á neðri hæð með straumnum; niður með á* 'below', *austur; austurátt* 'east', *langt (í burtu); fjarri* 'away', *hér í kring; hér nærri* 'around', *inn; innanhúss; inni; inni í landi; inn í land* 'inside' *nærri landi; að landi inni; inn; innan; að innanverðu* 'on land', *á staðnum nálægur; nákominn; naumur; nálægt* 'nearby', *norður; norðurátt* 'north' *hvergi úti; utan dyra; undir beru lofti, úti; út; frammi; utanhúss útbyrðis* 'outside', *fyrir borð landleiðis; á landi erlendis; til útlanda; handan hafsins* 'abroad', *suður; suðurátt* 'south', *undir fæti; neðanjarðar; fyrir neðan; undir á fótinn* 'underground', *upp á við uppi; uppi á lofti; upp; á efri hæð* 'upwards', *andstreymis; upp ána* 'upstream', *vestur, vesturátt* 'west'.

3.2 Data for Lexical Variation

In regards to lexical variation between the registers, I looked at the raw frequency of specific lexical items in the corpora. Table 3.13 below shows the difference between the most frequent words in each corpora. The ten most frequent were the following:

Table 3.13

10 Most Frequent Words in Each Register

Newspapers			Blogs		
count	word	Gloss	Count	word	gloss
492	Í	In	513	Að	to (inf) or at (prep)
418	Að	to (inf) or at (prep)	386	Og	And
297	Og	And	341	Í	In
290	Á	On	325	Ég	I
170	Sem	that (relative)	266	Er	Is
141	Til	to (prep)	245	Á	On
126	Um	About	159	Það	it (or expletive 'there')
124	Er	Is/Am	129	Sem	that (relative)
98	Við	with (prep) or we	119	Ekki	Not
97	En	But	111	Var	Was

The differences are perhaps not very surprising at this point. A look at the freqency of the words that did not show up on each respective list is more telling in this sense.

Table 3.14 Newspaper-specific Most Frequent Words

Newpaper Words	Freq. In Newpapers	Place in Newpapers	Freq. In Blogs	Place in Blogs
Til	141	6	71	17
Um	126	7	62	21
Við	98	9	93	12
En	97	10	103	11

Thus, these differences are not very great. The most common word stock is essentially the same. The two most common ‘newspaper’ words that did not show up in the blog list are about half as frequent as in the newpapers. The other two are about as frequent, and it could be considered mostly a coincidence that they did not show up on the blog´s ‘top ten’. The blog list is much the same:

Table 3.15 Blog-specific Most Frequent Words

Blog Words	Freq. In Blogs	Place in Blogs	Freq. In Newpapers	Place in Newpapers
Ég	325	4	5	N/A
Það	159	7	50	21
Ekki	119	9	56	15
Var	111	10	93	11

Other than *ég* 'I', which we have already seen to be more common to blogs than newpapers for reasons explained above, we see much the same pattern.[11] The next two, *það* 'it' and *ekki* 'not', are about half as frequent in newspapers as in blogs. The last, *var* 'was', is close to the same in frequency, and was actually the 11th most common word in newspapers.

3.3 Data for Syntactic Variation

Eliminating those instances described in section 2.4, all subjectless clauses, whether passive or impersonal, have a choice between Expletive Insertion and Topicalization.[12] The distribution of these structures is shown below in Table 3.17.

Table 3.16 Topicalization vs. Expletive Insertion

	Blogs	Newspapers	Total
Expletive	45	8	53
Topicalization	26	106	132
Total	71	114	185

This distribution is significant (Chi-square=67.99 p < .001 d.f.=1

Although newspapers had more impersonal sentences, it almost always preferred Topicalization to Expletive Insertion, and blogs have considerably more expletives than newspapers. I return to these numbers below.

[11] See the data in section 3.1 and the discussion in section 4.1.

[12] Here, in the vein of Rögnvaldsson and Thráinsson (1990), I ignore any distinction between Stylistic Fronting and Topicalization. This is because either process serves the same function in these types of clauses. It is in relative clauses where I will make the distinction and use the term Stylistic Fronting, since SF can take place in relative clauses and Topicalization cannot.

I also considered relative clauses, where there is another variable choice. This time, though, the choice is whether to undergo Stylistic Fronting or not. If not, the subject gap described in section 1.4 is left open, and the word order is just as in English. If the clause uses SF, some constituent is fronted to the pre-verbal position, or, as assumed above, to SPEC, TP.[13] There are many clauses, though, where SF cannot take place simply because there is either (1) an overt subject, thus blocking such movement, or (2) nothing in the clause which could be fronted. I eliminated all such clauses from consideration and only considered those clauses where SF was grammatically possible. The results are shown in Table 3.18.

Table 3.17 Stylistic Fronting

	Stylistic Fronting		
	Blogs	**Newspapers**	**Total**
Non-SF	17 (=65.4%)	17 (=30.4%)	34
SF	9 (=34.6%)	39 (=69.6%)	48
Total	26	56	82

This distribution is significant (Chi-square=8.98 p < .01 d.f.=1

Once again, newspapers seem to prefer movement and blogs seem to prefer constituents to stay put, all other things being equal.

4 Results & Interpretation

This section presents the results and interpretation of the previously shown data. Section 4.1 is concerned with the Dimension 1 features, section 4.2 with the lexical

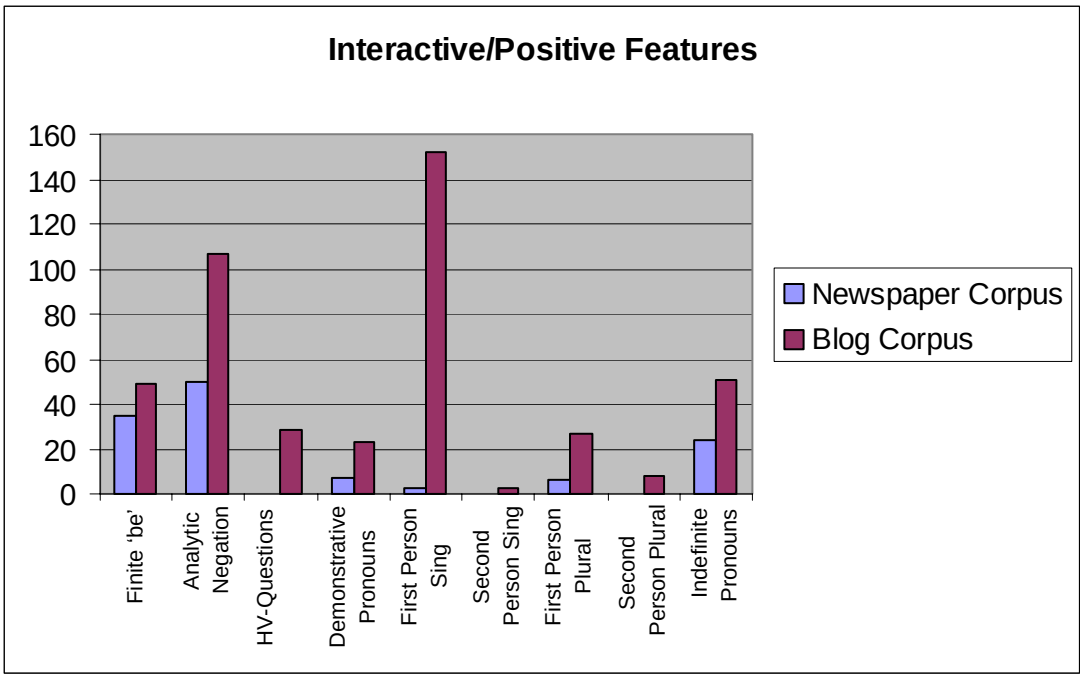
[13] There is a very strict hierarchy in regards to what constituents, when available, can raise in SF. Maling (1990) shows this to be *ekki* 'not' > predicate adjective > {either past participle or verbal particle}. The details of this are not important here, but it is important that these movements are not 'free.' Rather, they follow very particular rules regarding what is allowed to be fronted given multiple possibilities.

variation and section 4.3 with syntactic variation. Section 4.4 discusses some overall conclusions of the study and register variation in general.

4.1 Results & Interpretation for Dimension 1 Features

The graphs in Figures 4.1, 4.2, 4.3, and 4.4 summarize the tables shown in section 3. Each feature was divided and weighed based on the number of instances vs. the logical possibility of instances. The result was a decimal with which we could easily compare the presence of these features in each register. Figure 4.1 shows the interactive features and Figures 4.2, 4.3, and 4.4 show informational features, which were divided up due to the fact that two features (mean word length and type-token ratio) could not be counted in the same kinds of terms as the other four (instances per opportunities).

Figure 4.1



What is immediately striking about the data is the high representation of these features in blogs as compared to newspapers. This confirms one of the main predictions of this study: even though we are dealing with Icelandic, these features are distributed in Icelandic in the same way that English studies have led us to believe that they would be. Similar results are shown and discussed below in Figures 4.2, 4.3, and 4.4. What is especially interesting about the above graph is the number of first-person vs. second-person personal pronouns. This supports our intuition about the genre. It is a highly interactive genre, but the interaction is one-way. The writer knows that his readers know him, and he is writing to them and for their benefit, but he is usually not overly involved with them, as much as they are with him. The goal of the writing is for the writer to talk about himself, and for other people to read about him; a high degree of two-way involvement with the readers is not necessary. Biber (1988) describes first-person pronouns as "markers of ego-involvement in a text. They indicate an interpersonal focus and a generally involved style" (pp. 225). This, then, supports our intuition about both the interactive and egocentric nature of blogs. Further, when second person pronouns are used, they are far more often plural than singular (25 vs. 9), since the writer is typically addressing his "reader base" rather than one particular reader. Online newspapers, on the other hand, have very few first- or second-person personal pronouns, showing the lack of involvement characterizing the genre; it is a genre meant to be objective, informational, and non-interactive. At no point in the corpus does the writing address the reader directly, as shown by the complete lack of any second-person pronouns. Similarly, the only time first-person pronouns make an appearance is in direct quotations, where it is important to get a first-hand source to strengthen the validity of the reporting.

We see similar results for other features that co-occur with first- and second-person personal pronouns in interactive texts: every feature is better represented in the blog corpus than in the newspaper corpus. Some of these are far less intuitively interactive, but well-documented as such in English texts. It does not seem obvious that finite forms of the verb *to be* would have anything to do with the level of interaction between an author and a reader in a register, but empirical analyses show that this is the case in English; such seems to be the case in Icelandic as well. The same goes for analytic negation. There seems to be little intuitive reason to imagine that analytic negation as opposed to synthetic negation would be more interactive, but this seems to be case. Oddly enough, synthetic negation is neither a positive nor a negative marker of register on this continuum. That is, an interactive text may or may not have a significant presence of synthetic negation, but it doesn't seem to be relevant. What is relevant is that in interactive texts, there will tend to be a higher amount of analytic negation than in less interactive, more informational texts; this appears to hold true for Icelandic just as in English.

HV-Questions are either interactive or rhetorical in nature. The rhetorical ones, of course, are not used with the expectation of a response; they are answered immediately by the author, such as in the following excerpt:

*"ég var að koma heim úr world class og tók lyftuna eins og vanalega upp í íbúðina okkar. **En hvað gerist**? Ég get nú aldeilis sagt **ykkur** það, hún stoppaði svona mitt á milli 1. og 2. hæðar og sat bara föst þar, alveg sama hvað ég ýtt á marga takka og þó ég hoppaði nokkrum sinnum (ég veit svosem ekki hverju ég hélt að það myndi bjarga!)."*

*"I was coming home from world class and took the elevator like usually up to our apartment. **But what happens?** I can certainly tell **you (pl)** what, it stopped right between the 1st and 2nd floors and just sat still there, no matter how much I pushed all the buttons and jumped several times (I don't know exactly who I thought that would help!)."*

Sometimes hv-questions are used to address the reader-base directly, in the form of a sort of trivia game, with the expectation of a response. In all cases, though, hv-questions, even when used rhetorically, address and acknowledge the reader base directly, and are usually accompanied by some nearby second-person pronoun, as is shown in boldface in the above excerpt.

The negative, or informational, features also showed up as expected, with a much more prominent presence in the newspapers than in the blogs. These features are described as having a much more informational purpose and are predicted to be more prominent when interactive features are less so. The graph below in Figure 4.2, using the same counting methods, scale, and color scheme to summarize the data presented in section 3, shows the stark contrast between the presence of interactive and informational features here. Two features, Mean Word Length and Type-Token Ratio, could not be counted in terms of instances per opportunities. The graphs in Figures 4.3 and 4.4 show the raw numbers for those features.

Figure 4.2 Informational Features

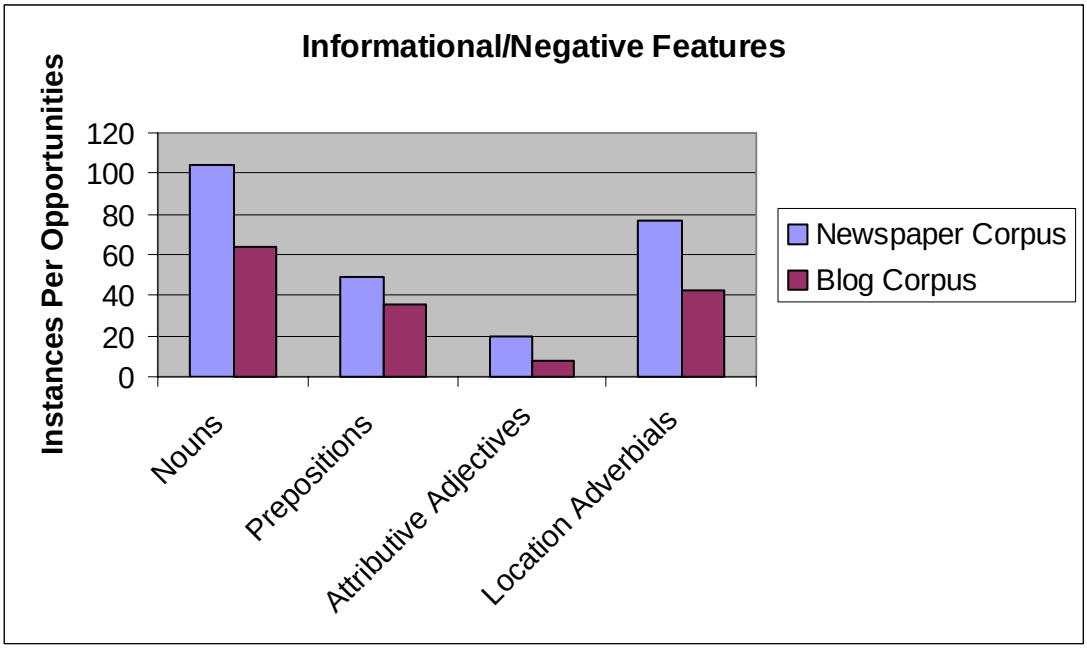


Figure 4.3 Mean Word Length

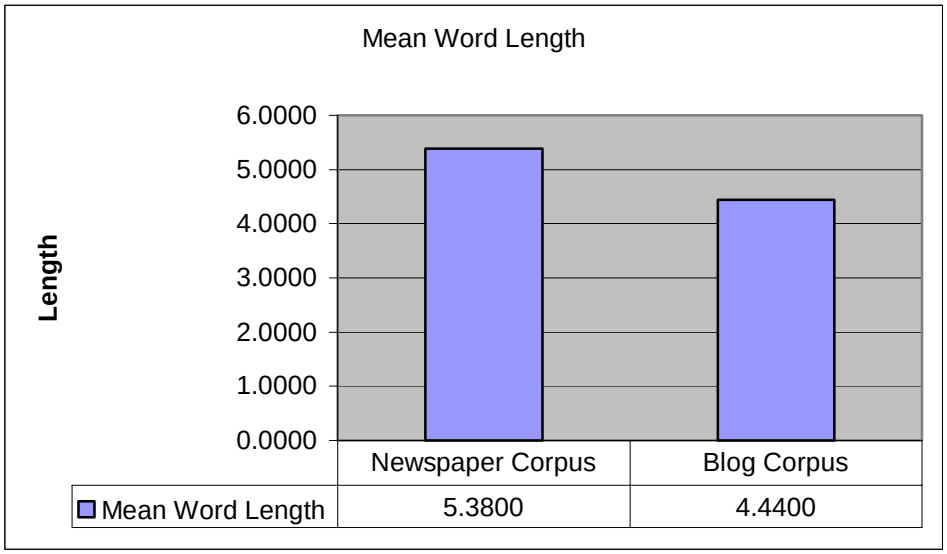
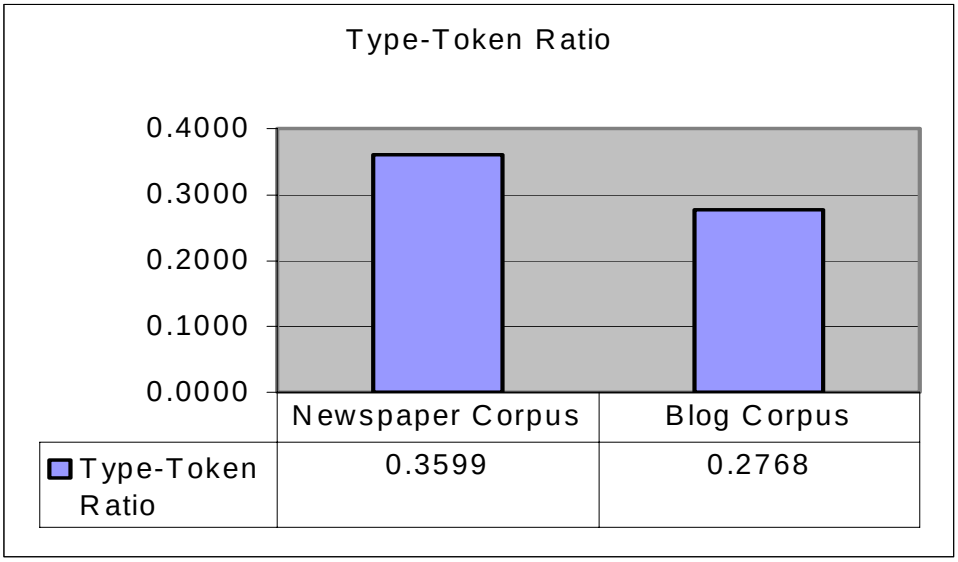


Figure 4.4 Type-Token Ratio



Since possibly the largest and most obvious contrasting element between the online newspaper and blog corpora was the noun phrases regarding person (see pronouns above), I looked into what kind of nouns occurred as subjects in main and subordinate clauses. I narrowed my search down to all verb-complement subordinate clauses (thus ignoring adjective- and noun-complement subordinate clauses, relative clauses, etc.) and looked at half of each corpus (5,139 words/Newspaper Corpus and 5,088 words/Blog Corpus). To this end, I looked at all nouns, but found no significant difference between singular and plural, or pronoun and non-pronoun (particularly, such as in the third-person), but an interesting difference in person. Unsurprisingly, almost all the newspaper subjects were third person, both in main and subordinate clauses. Blogs, however, were more interesting. Out of all the first-person subjects, almost two-thirds were in main clauses, and the opposite was true of third-person subjects:

Table 4.1

Subject in blogs	Main Clause	Subordinate Clause
1st Person Subject	25	13
3rd Person Subject	16	26

This could be interpreted as an effect of "who is talking about whom," the first-person (blogger) giving opinions about other third-persons (third-parties), and the third-person not nearly so often giving opinions about the writer. This isn't always the case, though, but it can be seen as a general trend. The third-person does, however, often give opinions about another third-person (as we will see below), and since any logical combination is likely to take place, the data are not necessarily interpretable in this way, at least not entirely.

Some examples of the first-person talking about the third person:

- Ég gat ómögulega skilið hvað þeir eru að gera þarna skokkandi á undan …
  I couldn't possibly understand what they are doing trotting there below…

- Sjáum hvað fréttirnar segja í kvöld.
  We'll see what the news says tonight.

- veit ekki hvað þetta er.
  (I) don't know what this is.

- ég held að um 1/3 af frammistöðu þeirra eru þessar frábæru sendingar og spil.
  I think that about 1/3 of their performance are these wonderful passes and games.

And some examples of the third-person talking about the third-person:

- það voru flestir sem sögðu að þeir ættu ekki trampólín af því þau voru alltof dýr.
  It was most that said that they didn't have a trampoline because they were far too expensive.

- Iceland Express auglýst í dag í Morgunblaðinu að þeir ætli að hefja flug til Frankfurt og verða fyrstu 1.000 sætin á sérstöku tilboðsverði eða 6.995,- krónur sem er náttúrulega hlægilegt verð
  Iceland Express adverstised today in the Morning Paper that they planned to begin flights to Frankfurt and the first 1,000 seats will be offered at the special sale price of 6,995 crowns, which is naturally a laughable price.

4.2 Results & Interpretation for Lexical Variation

The most common shared words were almost all function words: conjunction *og* 'and'; and prepositions *í* 'in', *að* 'at' (or an infinitive marker or complementizer), *á* 'on'; and the relativizer *sem* 'that'. The other was *er* 'is/am'. The register-specific words tend to be telling of the register they come from. Three out of the four most common newspaper-specific words were prepositions *til* 'to', *um* 'about', and *við* 'with'.[14] The other, *en* 'but', should perhaps not even be considered a newspaper-specific word since it is the eleventh most common word in blogs. Prepositions, of course, have been shown to be a feature more common to newspapers (as an informational register) than to blogs (as an interactive register). The fact that all three of the most common words to newspapers that were significantly less common in blogs were prepositions, then, helps to confirm prepositions as a feature of informational non-interactive registers.

A similar pattern is seen the blog-specific words. The most common of them is a personal pronoun, *ég* 'I'. This, of course, is clearly a feature of blogs as an interactive register. The next most common, *það,* is either a third-person neuter pronoun 'it' or an expletive. Although third-person pronouns are not necessarily considered to be a feature of interactive texts, expletive use, as will be discussed below (in section 4.3), is. Thus, *það*'s function as a pronoun combined with its function as an expletive makes it three times more prevalent in blogs than in newspapers. The third most common blog-specific word was *ekki* 'not', which, of course, signifies analytic negation, an attested interactive feature. It is more than twice as common in blogs as in newspapers. The last word, *var*

[14] *Við* is also a first-person plural personal pronoun 'we'. But, as shown earlier, this use of the word was not very prevelant in newspapers. The fact that it was one of the most common words in the newspaper corpus is undoubtedly due to its function as a preposition.

'was', like the last word on the newspaper-specific list, was the eleventh most common word in newspapers, and should thus not be considered a blog-specific word.

4.3 Results & Interpretation for Syntactic Variation

Figures 4.5 and 4.6 are pie graphs representing the distributions of Stylistic Fronting in blogs and newspapers. As explained in section 3, the percentages are based on the number of times where SF was possible in relative clauses; those clauses where the subject position was already occupied or where there was nothing to front were excluded.

Figure 4.5 Stylistic Fronting in Blogs

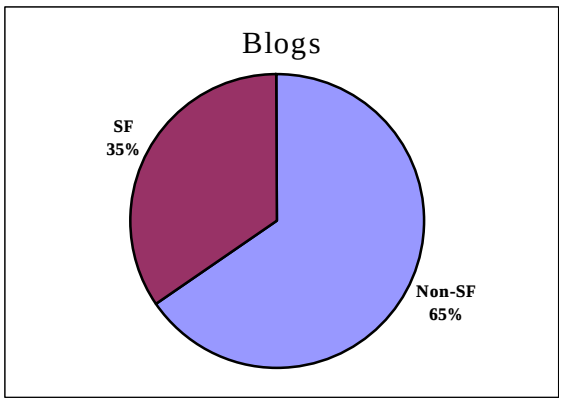
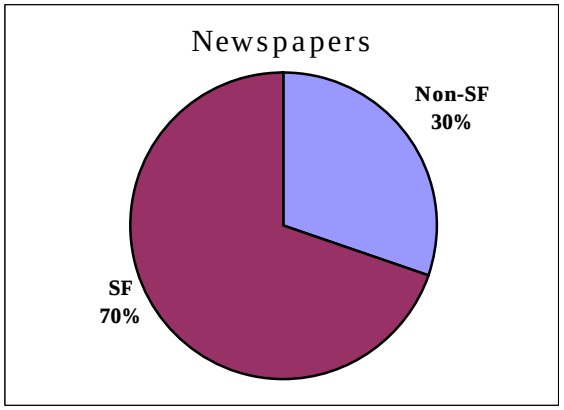


Figure 4.6 Stylistic Fronting in Newspapers

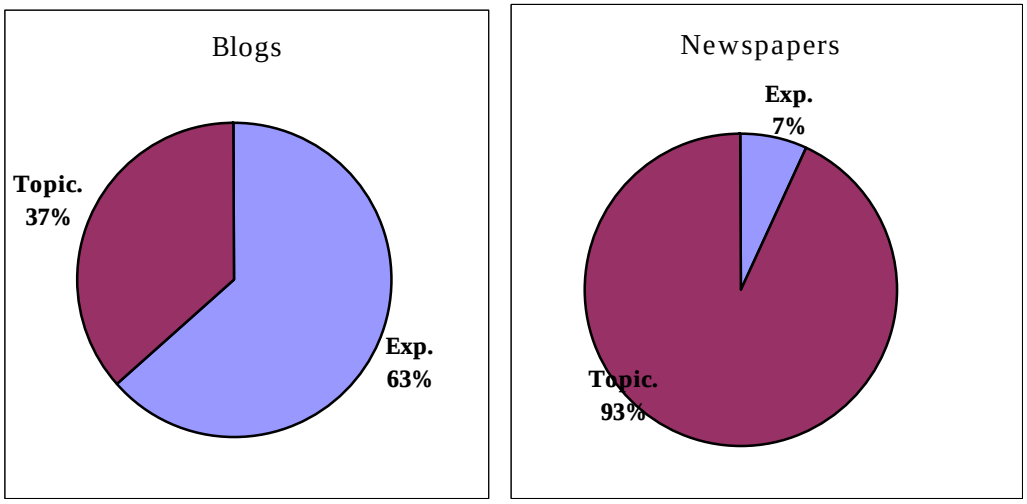


The difference here is fairly clear. Newspaper writers seem to prefer the construction where a movement operation fills the subject gap, whereas blog writers seem to prefer to avoid movement, leaving the subject gap unfilled. This might be motivated along the same continuum we have been discussing. That is, that Stylistic Fronting is an informative linguistic feature, co-occurring with the other informative features; thus, lack of prominence of such a feature is common in more interactive texts, in this case blogs. If

this is the case, we would predict that academic prose would have an even higher percentage of SF and informal conversations even less.

Or, this could be one instance of a more generalized trend in register variation, namely, preferring movement operations in more informational (or along another continuum, such as formality) texts. This seems to be the case when we look at the other syntactic construction we are concerned with, topicalization in impersonal main clauses. Figures 4.7 and 4.8 show distributions of the choice between Topicalization and Expletive Insertion in impersonal main clauses.

Figure 4.7 Impersonal Main Clauses – Blogs **Figure 4.8 – Impersonal Main Clauses - Newspapers**



The distribution in the blogs is almost identical, again preferring to avoid movement but in this case generating an expletive in the specifier of the tense phrase in order to do so. The newspapers, however, show a difference, preferring movement even more in impersonal main clauses than in relative clauses. This might be due to there being something especially marked about expletives in more informational (or, perhaps formal) registers. That is, since movement can do the job (filling SPEC,TP), generating an essentially meaningless word in a register where an emphasis seems to be on compact

delivery of information appears to be quite unusual. This seems even more to be the case when we look at the instances where expletives do appear. Out of the eight impersonal main clauses in which expletives were chosen over movement, five were quotes and one was a quotational subordinate clause. As for the other two, one was a political article for which I can offer no explanation, and the other is surrounded by evidence of a loosening of formality in a feature-type article (expletive is bolded):

- *Grín eða alvara? Íslenska eða pólska? Gildir einu, því* ***það*** *vantaði nauðsynlega tvo menn í vinnu hjá glerfyrirtækinu Íspan...*

- Joke or serious? Icelandic or Polish? Doesn't matter, since exactly two men were need (lit. **it** was needed two men) to work for the window company Ispan...

Right before the expletive is a case of two fragmented phrasal questions, and the 'expletive-dropped' verb phrase *gildir einu* 'doesn't matter'.[15] Note also, as discussed above, that many of the cases of subject dropping are clauses where the writer chose not to generate an expletive, showing again that expletives are marked in this register.

4.4 Overall Results and Interpretation

It has been shown in the multidimensional studies and in this study that many of the featural distributions which characterize a situationally defined variety do so unintuitively. That is, although features such as first-person pronouns are obvious in their level of interactivity, features such as analytic negation are not as obvious. I would like to

[15] I say 'expletive-dropped' because this verb phrase, *gildir einu*, translates literally to something like 'obtains one' or 'chooses one'. That is, it is conjugated in the third-person singular and has no subject. Thus, I expect that some 'expletive' was not generated or dropped for expositional puposes. Subject dropping of any kind, as shown above, does not tend to be a feature of an informational register such as newspapers. Its appearance here can be seen as some sort of loosening of standards which is certainly related to the two phrasal questions and the expletive under discussion here.

hypothesize as to why this is so; that is, why registers seem to vary so consistently on levels its users could not have possibly intended: registers are acquired tacitly, from exposure, just like like languages and dialects.

When a native speaker of a language acquires a new language variety or register, he acquires the properties of that register at all levels of the language. Such variation carries associative meaning, so that it can be recognized by other speakers of that variety. It is in some ways similar to language or dialect acquisition. First, it is learned tacitly by exposure. Like language acquisition, some levels of it may be taught. But most of the properties of that variety are learned at the sub-conscious level. The crucial difference between acquiring a dialect/language and a situational register is that the former may contain new structures and properties, which may be unrelated to another language/dialect, while the latter is produced by the same grammar as the language, but exploits that grammar in different ways.

What is important here is that we view a register as a language variety that is readily recognizable by a competent speaker who may not know anything about the properties that define that variety. This applies to all levels of language production or comprehension (an important distinction which I will discuss below), and applies especially in situations where referential meaning can be realized by two or more relatively equivalent variants. Take, for example, the choice a speaker has between:

a. The pen I write with

b. The pen with which I write

Both refer to the same object and give the same functional information about that object. The difference between the two is most easily defined, in a vacuum, as one of formality,

(b) being more formal than (a). In a vacuum, though, this is a single construction that can be taught. A teacher could say (as has historically happened), “Don’t end a sentence with a preposition. Use (b).”

But registers have bundles and bundles of linguistic variables and features, most of which cannot be taught in this way, at least not efficiently. It has been shown, for example, that more informationally-driven registers will feature more complex noun phrases. This is not taught, but it is acquired in similar ways as language. Consider the two variants below:

c. The informationally-driven registers

d. The registers where being informative is important.

It is hard to decide which is more complicated. (c) has the most compact delivery of information, a single noun phrase, but (d) has simpler structure in each phrase. That is, the noun phrase in (c) consists of four constituents. The informational consequences of such a compact structure may take more information processing: we have to find the head, categorize the modifiers (as determiner, adverb, adjective, in this case) and decide what is modifying what. But (d) has more structure: there is an NP, a CP, two more NPs (one which modifies the other), a VP and a predicative AdjP. Any speaker of English who has had enough exposure to various registers would recognize a difference in appropriateness between the two, without necessarily recognizing the linguistic difference.

The difference has to do with an overall schema of rule-governed variation. Different varieties prefer different structures. To keep to the example above, let’s imagine that (c) is more common in newspapers and (d) is more common in blogs. We could

postulate a Variable Rule which has various application rates depending on the situation, all other things being equal. In this case, tacit grammar would have a variable rule something like (e):

e. Variable Rule: Combine as much information in any given phrase as possible.

I formulate this rule in the sense of production, but this is not necessary. If a newspaper reader encountered a structure such as (d), he would feel as strange about it as a competent reporter would producing it. Rule (e) would have a higher application rate in newspapers than in blogs. Complex NPs, though, were not part of the present study. But we can summarize Topicalization with a Variable Rule such as (f):

f. Variable Rule: Avoid filling subject gaps with expletives.

Rule F would have an application rate of 37 percent in blogs and almost 100 percent in newspapers. (The exceptions were described above; almost all of them were quotational. Quotations probably have a different application rate due to different linguistic circumstances.) We could propose a similar rule to account for stylistic fronting. The overall idea, though, is that these rules carry situational meaning which is interpretable on both the producing side and the comprehending side.

This is why a stuffy academic might be laughed at for speaking too academically in informal conversations; academic language sounds silly in conversational settings (unless, of course, the topic is of an academic nature). The academic, in this case, is either less competent in conversational registers or has decided, for whatever reason, not to use those registers. Notice that the ability to recognize inappropriately used academic language does not require any special knowledge on the part of the listener; it requires

only that the listener hear something that seems linguistically inappropriate, though not grammatically incorrect: tacit knowledge.

These rules, then, do many of the same things that dialect does. They index exposure to a variety and therefore participation in a community, even among speakers who have never met. The varieties a speaker commands can be at the passive level (the ability, within a vacuum, to understand a string of text to be a newspaper article or a blog discussing the same thing, but not necessarily the ability to competently produce either) or the active level (production as well as recognition). This is another case similar to language/dialect: I may understand Italian/New England Working-Class English and recognize it immediately for what it is, but if asked to produce it, the output would be significantly different from a competent speaker. On the other hand, if I'm bilingual or bidialectal (or bivarietal), my competency may extend to production as well.

Linguistic tendencies toward higher and lower application rates of these variable rules exist in registers. (I stress again the definition of *register* as a variety that would be recognizable by any speaker of a language, with or without any specific language training). The enterprise of examining such tendencies could tell us how language, dialect, and situationally-dependant varieties are stored in the mind (such as the pragmatic functions of linguistic features) and the greater context of Universal Grammar, as well as give us special tools for the teaching of a language to non-native speakers. Further, we could teach native speakers why a certain variable construction is preferable to another, which would be especially useful in tutoring writing to both non-native speakers of a language and native speakers who have not acquired a specific variety, such as Academic or Business language.

5 Conclusion

The findings in this study show that some linguistic features do seem to carry their functions across linguistic boundaries. All twelve of the primary features in this study behaved as predicted in each register, and these predictions were based on a previous study of English. Furthermore, the distributions of some syntactic constructions seem to follow these generalizations as well. The extent to which they co-occur on a larger, more statistically significant scale, though, remains to be seen. Constructions such as Stylistic Fronting and Topicalization could be seen as formulations of variety tendencies, as such that they involve leftward movement and a more compressed delivery of information. The fact that they appear more often in newspapers than in blogs seems to support this idea.

Cross-linguistic register variation reveals a lot about human language in use and the extent to which the pragmatic uses of language could be universal. Biber's studies have gone a long way in showing this, but in order to learn how this holds up, more languages must be analyzed. Icelandic would make a strategic next choice for a multidimensional study for several reasons. First, it is similar in many ways to English, both structurally and functionally. Inasmuch as it is similar, the use of a different language could shed some light on which linguistic featural functions are language dependant and which are culture dependant. Second, it differs from English in some important ways. This is especially true concerning the very strictly rule-governed but considerably freer word orders allowed. Studying structures such as Topicalization and Stylistic Fronting, along with other language-unique constructions, could tell us a lot about language universals or universal tendencies from the perspective of register. Third,

it has a very wide array of uses and registers. Iceland is well enough developed that registers could be collected from almost every conceivable source, from hand written letters to advertisements, movies, or news broadcasts. Its literary history goes back almost as far as English's, lending itself nicely to potential diachronic multidimensional study. Lastly, it is geographically isolated and linguistically consistent. There is very little dialectal variation, allowing a much more controlled study which could look at situational variation without having to worry about interference from dialectal variation, as would be the case with other Germanic languages similar to English such as German or Swedish. Eventually, the extensive study of Icelandic register would reveal a great deal about what kinds of situational variation are language-dependent, what kinds are culture-dependent and what kinds are universal.

References

Adger, David and Smith, Jennifer. (2005). Variation and the minimalist program. In Leonie Cornips and Karen P. Corrigan (Eds.) *Syntax and Variation: Reconciling the Biological and the Social* (pp. 149-178). John Benjamins Publishing Company: Philadelphia.

Atkinson, Dwight, and Biber, Douglas. (1994). Register: A Review of Empirical Research. In Douglas Biber and Edward Finegan (Eds.) *Sociolinguistic Perspectives on Register* (pp. 351-385). Oxford University Press: New York.

Barðal, Jóhanna. (2003). Morphological case, syntactic functions and thematic roles in Icelandic. In Jorunn Hetland and Valéria Molnár (Eds.), *Structures of Focus and Grammatical Relations* (pp. 149-186). Tubingen: M. Niemeyer.

Biber, Douglas. (1988).*Variation Across Speech and Writing*. Cambridge University Press: New York.

Biber, Douglas. (1995). *Dimensions of Register Variation.* Cambridge University Press: New York.

Biber, Douglas and Finegan, Edward (Eds). (1994). *Sociolinguistic Perspectives on Register.* Oxford University Press: New York.

Bybee, Joan. (1998). A Functionalist Approach to Grammar and its Evolution. *Evolution of Communication, 2*(1), 249-278.

Connor, Ulla and Upton, Thomas. (2003). Linguistic Dimensions of Direct Mail Letters. In Pepi Leistyna and Charles F. Meyer (Eds.) *Corpus Analysis: Language Structure and Language Use* (pp. 71-86). Rodopi: New York.

Crystal, David. (2001). *Language and the Internet.* Cambridge University Press: New York.

Davies, Mark. (2006, Forthcoming). Towards the first comprehensive survey of register

variation in Spanish. In Eileen Fitzpatrick (Ed.) *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse.* Rodopi: New York.

Edmondson, Jerold A. and Donald A. Burquest. (1998). *A Survey of Linguistic Theories*. Summer Institute of Linguistics: Dallas.

Ferguson, Charles A. (1983). Sports Announcer Talk: Syntactic Aspects of Register Variation.'' *Language in Society, (12),* 153-172.

Holmberg, A. (2000). Scandinavian Stylistic Fronting: How any category can become an expletive. *Linguistic Inquiry, 31,* 445-483.

Hrafnbjargarson, Gunnar Hrafn. (2004). Stylistic Fronting. *Studia Linguistica, 58*(2), 88-134.

Labov, William. (1972). *Sociolinguistic Patterns.* Philadelphia: University of Pennsilvania Press.

Mardh, Ingrid. (1980). Headlinese: On the Grammar of English Front Page Headlines. *Lund Studies in English 58*. Lund: CWK Gleerup.

Maling, Joan. (1990). Inversion in Embedded Clauses in Modern Icelandic. In Joan Maling and Annie Zaenen (Eds.) *Syntax and Semantics: Volume 24, Modern Icelandic Syntax* (pp. 71-90). Academic Press: New York.

Rögnvaldsson, Eiríkur and Thráinsson, Höskuldur. (1990). On Icelandic Word Order Once More. In Joan Maling and Annie Zaenen (Eds.) *Syntax and Semantics: Volume 24, Modern Icelandic Syntax* (pp. 3-38). Academic Press: New York.