2015-12-01

A Corpus-Based Analysis of Russian Word Order Patterns

Stephanie Kay Billings
Brigham Young University - Provo

A Corpus-Based Analysis of Russian Word Order Patterns

Stephanie Kay Billings

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Deryle Lonsdale, Chair
Mark Davies
Jennifer Bown

Department of Linguistics and English Language

Brigham Young University

December 2015

ABSTRACT

A Corpus-Based Analysis of Russian Word Order Patterns

Stephanie Kay Billings
Department of Linguistics and English Language, BYU
Master of Arts

Some scholars say that Russian syntax has free word order. However, other researchers claim that the basic word order of Russian is Subject, Verb, Object (SVO). Some researchers also assert that the use of different word orders may be influenced by various factors, including positions of discourse topic and focus, and register (spoken, fiction, academic, non-academic). In addition, corpora have been shown to be useful tools in gathering empirical linguistic data, and modern advances in computing have made corpora freely available and their use widespread. The Russian National Corpus is a large corpus of Russian that is widely used and well suited to syntactic research.

This thesis aims to answer three research questions: 1) If all six word orders in Russian are possible, what frequencies of each order will I find in a data sample from the Russian National Corpus? 2) Do the positions of discourse topic and focus influence word order variations? 3) Does register (spoken, fiction, academic, non-academic) influence word order variations?

A sample of 500 transitive sentences was gathered from the Russian National Corpus and each one was analyzed for its word order, discourse pattern, and register. Results found that a majority of the sentences were SVO. Additionally, a majority of the sample contained the topic before the focus, and most of the sample were from the non-academic register. A chi-square analysis for each research question showed statistically significant results. This indicates that the results were not a product of chance, and that discourse patterns and register influence word order variations. These findings provide evidence that there is a predominant word order in Russian.

Keywords: Russian, word order, corpus, discourse analysis

ACKNOWLEDGEMENTS

Many people were vital to my efforts to complete this thesis. I appreciate my committee for their willingness to help me on a tight schedule: Dr. Jennifer Bown for her Russian language insights, Dr. Mark Davies for his help in devising the methodology and for his expertise in corpus linguistics, and Dr. Deryle Lonsdale for patiently and encouragingly ensuring that my thesis was ready for defense. I also recognize Dr. Heather Sturman for her patience and help through the beginning stages of this thesis, and LoriAnne Spear for coordinating and facilitating all of the logistics.

My support system outside of the university deserves many thanks. I am grateful for the sacrifices from Irene Beyrich, Tiffany Collette, Callan Olive, Sarah Cox, Melissa Billings, Sharon Billings, and Hannah Scott who took care of my son while I was in class or needed extra time in my schedule for research and writing. All my family and friends who believed in me and encouraged me to finish deserve thanks as well. Elise Thompson and Dr. Mindy Ly found me the right diagnosis after my years of illness, and I cannot thank them enough for it. And my son Charlie deserves a standing ovation for his years of patience with me.

My husband, Greg Billings, made all of this possible in every sense. Through these years of my graduate work he has been as committed as I have been (and sometimes more so) to seeing this dream to fruition, and has never complained. Every good thing I have accomplished in the years since we met is due to his unfailingly patient and fiercely loyal love. Thank you so much, Greg. I love you forever.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Some scholars claim that Russian has free word order, meaning that the subject, verb, and object can appear in any order in a sentence. For example, Dyakonova (2009) asserts, "A notorious property of Russian is its free word order, which is constrained by discourse factors." (p. 43). However, Russian language experts disagree about how free the word order actually is. Kallestinova (2007) maintains, "It is well known that Russian has a relatively free word order." (p. 1). Other scholars claim that Russian may have a somewhat fixed word order. They hypothesize that although Russian words can occur in any order in a sentence, some word orders are regarded as more correct and thus are more frequently used whereas the other word orders are more marginal and used less frequently. Unfortunately, researchers on both sides of the question don't have much empirical evidence to substantiate their claims. This leaves the issue of Russian's supposedly free word order unresolved.

Resolving this issue with more data is valuable to all kinds of language experts. For example, researchers could have more information about the syntax and typology of Russian, leading to more elegant typology criteria. Russian language instructors could help their students achieve more rapid fluency in the language by relating the word order patterns. Translators and interpreters could be able to communicate more efficiently by following the patterns shown in the data.

Researchers have collected some usage data in order to see whether the word order of Russian is free or somewhat fixed: some have collected speech samples from groups of native Russian speakers, and others have collected small samples of usage data from the internet or

literary archives. These samples lead researchers to posit that Russian speakers may indeed prefer to use some word orders more than others. The empirical samples are a good start, but they lack large amounts of data from various sources. Larger, more diversified samples are more representative of the actual language usage and provide better illustration to resolve the question of how free or fixed Russian word order is.

In this thesis I use a large sampling of diversified data from a freely available online database called The Russian National Corpus. I analyze the sentences from that sample to determine what word orders they exhibit, and try to better understand what factors may influence the word order patterns that I see in the data. I then employ statistical tests to show that the data I found is meaningful. Accordingly, I make reliable inferences about the word order patterns of Russian and add my conclusions about how free or fixed Russian word order is.

Chapter 2: Review of Literature

Russian employs 6 grammatical cases: nominative, accusative, genitive, prepositional, dative, and instrumental. The thematic roles of a sentence are marked with suffixes on all nouns and adjectives, and each suffix indicates case, gender, number, and person. The only exceptions to case marking are indeclinable nouns, which are mostly foreign borrowings. The morphological marking of the cases allows for freedom in the syntax; the pervasive use of the cases in Russian means that the word order does not play a role in signaling grammatical relations. Russian is described by some researchers as a language with free word order, meaning that six different word orders (SVO, SOV, OVS, VSO, OSV, and VOS) are possible. Upon reviewing the literature, it seems that many researchers agree that some word orders may be regarded as more acceptable to speakers than other orders are. This section will survey the literature on Russian word order, and talk about the important differences between formal data and empirical data and their usefulness in syntactic research. The importance of corpus research will also be discussed, as well as the usefulness of the Russian National Corpus and theoretical notions of topic and focus.

Brief Overview

The functional approach to word order theory laid important groundwork for the literature on Russian word order. The functional approach, which accounts for word order variations by analyzing discourse patterns, was largely developed by the Prague Circle as a means of investigating Slavic language. They developed the bipartite division of the sentence into theme and rheme, which correspond to given information and new information. Later Mathesius developed the term transition, which led Firbas to posit a tripartite division of the

Russian sentence into theme, transition, and rheme (Kallestinova, 2007). The tripartite division, or three-way sentence division, was further developed by other linguists, setting the stage for other investigations into the word order patterns in Russian.

Other researchers built on the work of the Prague Circle. For example, Kriesing (1977) dissected the contemporary literature of Russian word order, concluding that syntactic components follow a linear order that must be viewed as fixed on the bases of grammar, communication, and stylistics. More specifically, the surrounding context of a sentence both conditions the syntactic relations and contributes to the meaning of the sentence; therefore it is necessary to consider the context of a sentence when analyzing its word order (Vintseler, 1977). These studies, among many others, call into question the common assumption that word order in a given Russian sentence is completely free; they also add to the theory that context and discourse are factors in Russian word order variations.

More researchers afterwards have stated that Russian word order is a mechanism of discourse management. Frink (1984) conducted a discourse analysis study of Russian cohesion. As with many formal papers, Frink doesn't mention where his data came from, which may limit how much the data can be generalized. He agreed with the Prague School and found that in Russian new information is placed at the end of the sentence as the rheme; the sentences begin with the old information, or theme. He claimed that under the model given in Halliday (1967), Russian uses 5 types of sentences: reference, ellipsis, conjunction, lexical cohesion, and cohesive word order. Similarly, Lehmann (1982) performed an analysis of the theme-rheme division in Russian. His findings indicated that the distribution of Russian sentence parts is constrained by the speaker's intention.

Among other authors, Blekher (1995) asserts that the basic word order of Russian is SVO, and cites other researchers who agree. She studied data taken from 4 genres of Russian texts: colloquial, fairy tale, neutral belles-lettres, and scholarly writing. Each genre was taken from a separate sample of roughly 375 pages each. She found that the sentence elements are consistently arranged in a "given information before new information" order and that marked orders are only used paragraph-initially to signal a shift in the flow of discourse. Her analysis continued, showing which constraints were at work in determining word order, and ranking the importance of the constraints. Blekher concludes that the "given before new" information packaging constraint is more important than the "basic word order" constraint. She also gave her reasons for why the marked orders signal a shift in discourse: the marked word orders (in which the subject is after the object or adverbial) are more difficult for readers to process, therefore they draw the reader's attention to the change in topic. Frink and Blekher show from their research that Russian discourse follows the same pattern as English and many other languages—it typically satisfies the general tendency of "given before new" information in a sentence. Researchers agree that this tendency to place new information at the end of a sentence is an information processing function. It is said that the mind is better able to integrate new information into its representation of the sentence when the old information precedes the new (Blekher, 1995).

Although Russian follows the same discourse patterns as other languages, certain constructions may be subject to word order constraints. Namely, locative inversion is disallowed in transitive sentences (Mezhevich, 2001). Mezhevich tested the Restrictions on Locative Inversion from Bresnan (1994) and found that they apply to Russian. In fact, she says that locative inversion is only allowed in some intransitive sentences and some passivized transitives.

She posited that the interpretive function of locative inversion in Russian is associated with definiteness. Specifically, Mezhevich found that when locative inversion is preverbal it occurs with definite noun phrases (hereafter NPs), and when the inversion is postverbal it occurs with indefinite NPs. She concluded from her analysis that word order, stress, and context all have important roles in definiteness interpretation of NPs in Russian. A potential weakness in Mezhevich's research is that she analyzed only sentences from her own intuitions. Although this is a common practice in modern linguistic research, such practices make generalization difficult.

An analysis was performed with data from the Russian National Corpus to determine whether Russian word order is subject to weight effects (Kizach, 2012). In contrast to the discourse analysis studies mentioned so far, this study utilized a parsing approach to analyze the data for constituent weight. A sample of roughly 600 sentences was gathered from the corpus, and the analysis showed that heavy constituents follow light constituents in four different constructions that were found: Postverbal PPs, Double Object Construction, Adversity Impersonals, and SVO word order. This shows that Russian word order is subject to the constraints of different constructions, and illustrates a corpus methodology that will be useful to this thesis.

Many of the researchers discussed so far have utilized discourse analysis methods to illustrate that word order in Russian may be somewhat fixed, based on different syntactic and semantic factors. Many of the studies discussed so far are empirical in nature, but rely on small samples. This is problematic for research because it is difficult to generalize about how the language works based on such small sets of data. It is also possible that when researchers only utilize intuitional data, they may have some bias—they are more likely to produce idiosyncratic sentences from their minds that fit their theories. This is, again, problematic for a study that

claims to understand how the language at large works. The research that I've discussed so far has shaped my thesis research on Russian word order variations.

The questions of this thesis will attempt to fill in some of the gaps of the studies mentioned so far. First, if all six possible word orders are possible, how frequently would each form occur in the given data set? Second, would the positions of given and new information in the sentence influence the variations in word order frequency? Third, would register (spoken, fiction, academic, or non-academic) influence the variations in word order frequency? In the next section, I will discuss the usefulness of empirical and corpus research, especially in terms of answering my research questions.

Empirical vs Formal Studies

The research that I have reviewed so far is short on corpus methodology. Moreover, the scarcity of corpus studies is found throughout many fields of linguistics. McEnery & Wilson (2001), in their first chapter, explain the reasons for this scarcity over the past few decades and highlight some of the tension between more formal methods and more empirical methods, which I will briefly survey in this section. Understanding the history of these scarcities and tensions will help to demonstrate the relevance of my research questions and the ongoing need for Russian corpus research.

In *Syntactic Structures,* Chomsky (1957) outlined everything that was wrong, in his view, with the state of empirically-based research in linguistics. Up until that point, researchers would often collect as much data as possible about a language into a body of data, or corpus, in order to try and make predictions and model how the language worked. However, Chomsky pointed out that the size of the corpora that researchers were collecting at the time were too small to yield useful insights into research questions. The researchers couldn't make reasonable claims about

how the language supposedly worked based on such sparse data. Chomsky argued that it was better to consult a native speaker's intuitions for data, even if that native speaker is yourself, than to attempt to create and use corpora. *Syntactic Structures* was widely accepted, and as a consequence, most data-based and corpus linguistics in the United States stopped for a few decades. Many researchers adopted the formalist tradition of using Chomsky's data collection method, and many papers and theories were based on formalist data until the 1980s.

The renewed interest in corpus linguistics was sparked by at least four changes. The first is that earlier corpora were too small, but advances in computing have made large corpora (like the British National Corpus[1]) possible and freely available. The second change is that corpus data has been shown to lend valuable insights into the real world, especially natural language processing. The third change is that researchers have been moving away from binary judgments of grammaticality towards more nuanced descriptions of phenomena, which corpus linguistics is well-suited for. The fourth is that researchers were finding that the theories that generative/formal linguists had posited were built on flawed data; there needed to be a standard for the data that they were using, and publicly-available corpora fit this need.

Empirical studies and the use of annotated corpora to gather data are becoming increasingly popular in linguistic research. Large corpora that can be divided into registers—such as spoken, fiction, non-academic, and academic—are used to determine whether certain constructions in the given language are more frequent in some registers than in others. A researcher could also search for one construction in the entire corpus and get a large picture of the phenomenon. A large corpus can show interactions between grammatical and lexical

[1] http://www.natcorp.ox.ac.uk/

categories, such as which verbs occur more frequently with which verb tenses. A corpus can also show morphological phenomena, like which prefixes and suffixes are attached to which word roots most often. A corpus is an invaluable tool that illuminates morphosyntactic patterns in language data that I would not be able to see otherwise. In order to weigh in on the debate about the word order patterns of Russian, I need to gather and analyze empirical corpus-derived data.

Empirical and Corpus Studies of Russian

The literature discussed so far have highlighted many factors which possibly contribute to the phenomenon of different acceptable word orders in Russian. Even though researchers agree that various discourse and pragmatic considerations may constrain the possible word orders, the researchers disagree about which of the considerations are most important. They also have many different kinds of analyses that bring us to these conclusions. Many of these analyses utilize rationalist data, which has been shown to be insufficient to answer the questions of this thesis. This section will highlight some more recent empirical studies of Russian word order which provided potential design ideas for my research methodology.

Kallestinova (2007) elicited sentences and grammaticality judgments from 237 native Russian speakers and found that in transitive sentences the orders SVO, OVS, and SOV are preferred. She also found that VSO, VOS, and OSV are not produced but still regarded as acceptable, although perhaps marginal. Through the use of a data set from a large number of speakers, this study showed that in spoken Russian, specific word orders are preferred over others. However, this study focused on spoken Russian only, which may limit how much the data can be generalized to other registers of the language. Russian is a language that allows for the structural encoding of topic and focus (Dyakonova, 2004). The topic/focus dichotomy is generally regarded as interchangeable with given/new and theme/rheme; I will henceforth use

topic/focus for the sake of continuity. Dyakonova agrees with Kallestinova that verb-initial sentence orders are rare in Russian and that they may be subject to an intransitivity constraint. Dyakonova performed a language acquisition observation of two groups of children: one group acquiring English natively and one group acquiring Russian natively. Perhaps unsurprisingly, she found that the Russian-speaking children made wide use of the pragmatically marked word orders. From her analysis of the experimental data, Dyakonova also concluded that only subjects of intransitive verbs functioning as topic can be placed before the verb. She also concluded that pragmatic constraints on word order are learned in parallel to syntactic constraints. Both of these studies build on previous literature in demonstrating that context and discourse pragmatics are important in determining word order in Russian. Based on the experimental data from many speakers, both researchers theorize that verb-initial word orders are rare in Russian. These findings set the stage for my thesis.

Many researchers have used corpora in their research of Russian. For example, Hentschel (1992) used data from a small corpus to hypothesize about the associations between case assignment and word order. Based on this analysis, Hentschel found that case assignment is dependent on constituent linearization and that this dependence seems not to be semantically motivated. Similarly, Grenoble (1998) used data from several small corpora containing taped, spoken Russian. Her analysis found that word order is determined by pragmatic rather than syntactic factors.

More recently, Malamud (2002) conducted a discourse-style analysis from a corpus of literary texts in Russian and investigated the relationship between attentional structure of discourse and word order. She utilized a formal discourse analysis model based around Centering Theory (Brennan, Friedman, & Pollard, 1987). This theory provides an algorithmic definition of

topic which is equivalent to the widely accepted definition of theme. The semantic entities of the sentence are ranked according to their salience in that sentence. Malamud took 44 segments which each contained 2 or more sentences, and at least one of the sentences in each segment was scrambled. She performed at least 2 differently-ranked analyses of the corpus data and found that putting an object in sentence-initial position did not affect its discourse salience, but when the subject was postverbal both the subject and the verb were less salient. Malamud concluded that the attentional structure of discourse and word order are interdependent phenomena in Russian. This corpus study shows that discourse analysis is indeed an important tool for researchers to understand the word order variations in Russian.

Additionally, other linguistic phenomena can be investigated with the help of a corpus. Alekseyenko (2013) conducted a parallel Russian-English study, compiling data from *National Geographic* and the Russian magazine *The World of Animals* into a small parallel corpus to compare the two languages. As shown previously, corpus research lends itself elegantly to answer many kinds of morphosyntactic questions.

All of the studies cited so far have used data that the researchers gathered from various sources and were custom-built for their research. These studies have shown that empirical and corpus data are useful to the study of Russian. However the ideal for corpus research is a large, balanced, and publicly available corpus. Such a corpus can and should serve as a standard for data collection. In the next section I will relate more about the current state of Russian corpus linguistics.

The Russian National Corpus

Corpus linguistics is increasingly more viable and more popular. Many large corpora of English have been made publicly available in the last couple of decades, like the British National

Corpus and the Corpus of Contemporary American English[2]. Similar corpora for other languages have also been created, such as the Corpus do Portugues[3] for Portuguese, Corpus del Español[4] for Spanish, and ARTFL[5] for French. These languages, which have somewhat similar morphological properties, allow researchers to utilize the same methods to create these corpora. Researchers have created taggers, which enable computers to automatically parse the data and assign part-of-speech tags. These taggers are incredibly robust, and are able to sort through data much faster and more efficiently than human researchers. They are the key to annotating large corpora, yet they only work for certain languages with certain morphological properties.

Although large, publicly available annotated corpora should serve as the standard, no such corpus existed for Russian until the Russian National Corpus. Sharoff (2006) talks about this in detail, and although his paper is almost a decade old now, its main points are still relevant to my research. In the past few decades, some researchers attempted to make large corpora for Russian. In the 1970s Zasorina (1977) and her colleagues created a 1 million word corpus patterned after the Brown Corpus, but this corpus was never made publicly available. The Uppsala Corpus[6] was developed in the 1980s in Sweden, which is made of 1 million tokens from 600 samples: 300 fiction samples and 300 non-fiction. It is accessible via the internet, but as Sharoff points out, it is too small and doesn't cover enough registers to get an accurate account of the usage of the language as a whole. Another large problem is that the Uppsala Corpus is not morphologically annotated nor lemmatized. This makes the corpus very difficult to use, as it cannot be searched for specific grammatical phenomena; the corpus can only return raw data to

[2] http://corpus.byu.edu/coca/
[3] http://www.corpusdoportugues.org/
[4] http://www.corpusdelespanol.org/
[5] https://artfl-project.uchicago.edu/
[6] http://www.moderna.uu.se/slaviska/ryska/corpus/

researchers. This lack of large annotated corpora left researchers of Russian at a disadvantage—robust corpus data allows for stronger claims about how the language works.

Sharoff posits that the main reason for the lack of a large Russian corpus is that Russian is a highly inflectional language:

> For instance, an adjective inflects for case, gender and number, giving 36 basic adjectival categories in total, while a verb in addition to its own 14 basic categories has up to 4 participles, each of which declines for adjectival categories. This leads to thousands of separate tags that cannot be searched effectively. (p. 175)

The morphological complexity means that in order to make a corpus that is morphologically annotated and lemmatized, there will be a high degree of ambiguity. Frequent word forms can correspond to several lemmas. In addition, there is ambiguity between forms in the same lemma. Take the example of *knigi*[7], which is the singular form of "book" in the genitive case but also the plural form in the nominative or accusative case. Even more complex is Sharoff's example: "…*znakomoy*, which is the singular form in the genitive, dative, or prepositional case of either a noun or an adjective." (p. 176)

Strategies that are commonly used for other corpora do not solve the ambiguity problem for potential Russian corpora. Using traditional part-of-speech taggers based on statistical models does not help the situation, because they tend to be based on word order restrictions, and Russian does not have such restrictions. Fortunately, Russian corpus linguists have come up with a

[7] For Russian words, I will use Romanized spellings instead of using Cyrillic. I will use the BCN/PGCN conventions. BCN/PGCN conventions were developed without the need for special characters, allowing for easy readability and clarity.

system of partial tagging that minimizes a lot of ambiguity by checking for the context of the other words in the sentence. However, this system cannot eliminate the ambiguity.

Sharoff and his colleagues have gone on to create the Russian National Corpus (RNC), also sometimes called the BOKR (BOl'shoy Korpus Russkogo yazyka). As of the last update in January 2008, there were a total of 149,357,020 tokens in the RNC, making it the largest balanced corpus of Russian available. It contains spoken material gathered from spontaneous spoken Russian and the transcripts of Russian movies which totals 3.9% of the total tokens in the corpus. Additionally there are fiction texts, which total 39.7% of the corpus, further divided into ten genres including adventure, sci-fi/fantasy, and drama among others. The search interface gives the option to exclude or include any or all genres from the search results. The non-fiction texts make up 56% of the total corpus and are divided into many different types and genres including journalism, technical, academic (12.8% of the total corpus), official business, day-to-day life, advertising, and electronic communication, to name a few. Each has the option to be included or excluded from the search results.

There are two different interfaces to this corpus and both are powered by Yandex, the Russian search engine. One of the interfaces can be freely accessed via a web browser[8] and gives options for searching the RNC and some of its subcorpora, searching other corpora, searching just the internet, or doing a combined search. The other interface[9] has options to search in English or Russian. The RNC is morphologically annotated and lemmatized using the method of partial tagging that was discussed in the previous section. Because of the resulting ambiguity, the creators of the RNC have disambiguated about 5.5 million tokens by hand—this subcorpus is

[8] http://corpus.leeds.ac.uk/ruscorpora.html
[9] http://ruscorpora.ru/en/index.html

called the Disambiguated Corpus, or the Deeply Annotated Corpus. The researchers have created dependency trees to mark the syntactic relations for each of the 5.5 million tokens in the Disambiguated Corpus. This is quite a feat of research, and although the size of the Disambiguated Corpus does not compare to the large corpora of other languages, it is still an improvement over the other corpus options that are available for Russian.

Many researchers have used the RNC in the short years since its release, some of whom are listed here. For example, Apresjian (2013) used the RNC and COCA to study the differences between emotive phrases in Russian and English. The RNC was also utilized to study the 'nu' suffix/infix and the phenomenon of its being dropped (Nesset & Makarova, 2012). Furthermore, Gracheva (2013) used the RNC for her thesis on contrastive suffixes in Russian. Fiedosjuk (2010) used the RNC for a study that compared Russian attitudes towards German and Polish people. Finally, the Russian Dependency Treebank SynTagRus, which is a subcorpus of the RNC, was used for a psycholinguistic study of the use of relative clauses and how that use affects speakers' memory (Levy, Federenko, & Gibson, 2013). The wide variety of the research questions being answered by the corpus data lends more credibility to the corpus.

My thesis research aims to use quantitative data to weigh in on the debate about the potential word order restrictions in Russian. In order to accomplish this aim, I will use the Disambiguated Corpus of the RNC to answer the following questions:

1. If all six possible word orders are possible, how frequently would each form occur in the given data set from the corpus?
2. Does the position of discourse topic and focus influence the variations in word order frequency?

3. Does register (spoken, fiction, academic, or non-academic) influence the variations in word order frequency?

For reasons already discussed, the RNC is not as robust or well-balanced as other large corpora like COCA or the BNC. However, it is the most balanced corpus of Russian that is currently available, and it is robust enough to satisfactorily answer my research questions.

Notions of Topic and Focus

At this point, it is important to define the notions of topic and focus that will be used in my research. These definitions will be modeled after the work of King (1993), whose dissertation about Russian topic and focus was particularly helpful.

From the time that the Prague Circle defined their notions of discourse topic and focus for Slavic languages, many papers have been published that analyze topic and focus in the languages of the world. Many of them utilized the two-part division, which has often been called the Functional Sentence Perspective or Topic Focus Articulation. With this method, all parts of the sentence are accounted for under either topic or focus. Any information that is new to the discourse is the focus, and all non-focused material is considered the topic. King identified two problems with the two-part sentence division: 1) some sentences contain material that is not focused nor topicalized, and 2) the verb can be particularly difficult to assign to either focus or topic. She utilized a three-way division of the sentence into topic, discourse-neutral/transitional information, and focus from Firbas (1965). King claimed that this particular method accounts for the specific Russian data better than other methods. Therefore I will utilize the three-part sentence division proposed for Russian in the data analysis. King explained these terms in greater detail, which I now summarize.

Topic and Discourse-Neutral Information

How is a topic identified? According to Krylova & Khavronina (1988), topicalized material is defined as the items that are of immediate interest to both speakers. In addition, topics tend to be definite and are often pronominal. King also asserts that topics in Russian are always preverbal. Finally, King includes a discussion of the difference between an external topic and an internal topic, which is mentioned here merely to further refine the notion of topic for discourse analysis methods. An external topic is not an argument of the verb (although it can be coreferential with one) and an internal topic is an argument of the verb. Therefore it is possible to have multiple topics in one sentence. It is also possible to have a sentence without a topic. These criteria define my notion of topic for this thesis.

Here are shown some example sentences of topics from King (1993, p. 73) in which all topics are bracketed:

(1) *[Na stolye] stoyala lampa*.
on table stood lamp
'There was a lamp on the table-TOP.'

(2) *[Lampa] stoyala na stolye*,
lamp stood on table
'The lamp-TOP is on the/a table.'

(3) *[Rasskazov] [ya] prochitala mnogo* .
stories.GEN I read many
'Stories-TOP , I-TOP read many of.'

(4) *[Ivan], [ya] [ego] ne lyublyu*.
Ivan.NOM I him.ACC not like
'Ivan-TOP, I don't like (him).'

(5) *[Vchera] priyekhala mama*.
yesterday came mother
'Yesterday-TOP, mother came.'

Notice that all of the topics appear in initial position in these examples. Note the difference between (1) and (2). In (1) the topic *na stolye* is definite while the subject *lampa* is focused, and according to King, frequently interpreted as indefinite. However in (2) when the subject appears before the verb and the PP is after it, the subject is definite, and depending on context, the PP may or may not be definite. Sentence (3) shows the possibility of two topics in one sentence. Sentence (4) shows an external topic which is external to the verb but coreferential with an argument of the verb. Sentence (5) shows that the adverb is the topic because the period of time is of common concern to the speakers. The examples shown and discussed here illustrate many different kinds of topics that are available in Russian. However when categorizing the data for statistical analysis, I will categorize topic in general and not refine that category any further.

By this definition and discussion of how to categorize a topic, I am able to define what discourse-neutral information is. Discourse-neutral information is defined as material that is non-rhematic (i.e., not new to the discourse) but also not topicalized; verbs often fall into this category, as they often perform a transitional function between topic and focus. Therefore anything that is not thematic material nor rhematic material is categorized as discourse-neutral. In this thesis, the criteria for identifying discourse-neutral material simplifies the categorization of the elements of the sentence, i.e. discourse-neutral material will remain uncategorized and will be left out of the analysis.

In performing the discourse analysis of the corpus data that was gathered, I found the need to modify King's diagnostic criteria slightly. In sentences that King identifies as having two topics, similar to (3), I found that one of the preverbal nouns was always mentioned in the preceding sentence. I thus identified such nouns as topics and the other noun, which was not mentioned in the preceding sentence, was therefore new discourse material.

Focus

Focus is generally described as information that is new to the discourse and not presupposed; an item is presupposed when it is an implicit assumption relating to a previous utterance in the discourse. From the previous definitions, I can also assume that any material that is neither topic nor discourse-neutral is therefore focused. King asserts that the most common focus pattern is when the focused constituent is in final position. King's VOS example (1993, p. 74) demonstrates this principle:

(6) *Chitayet knigu [otyets]*
Read.3rd.sg book.ACC father.NOM
'Father-FOC is reading a book'

This shows that when the subject 'father' is the focus in (6), it is found in final position. There is an additional way to identify focus in a Russian sentence. The particle *zhe* is an intensifier and thus it lexically marks focus in a sentence. Note the following example from King:

(7) *On uyedyet [sevodnya zhe]*.
He depart.PERF today FOC
'He will leave today-FOC'

Although the focused constituent is often in the final position of the sentence, *zhe* can appear in any position in the sentence, and always marks focus. Researchers disagree about whether *zhe* is a clitic, and King (1993) merely states "If the focused element is greatly stressed, the emphatic particle *zhe* can appear on it…" (p. 158). Regardless of the status of *zhe,* it remains helpful in my criteria for determining focus.

There are three types of focus: new information focus, presentational focus, and contrastive focus. Contrastive focus picks out an element from a presupposed set of alternatives. The following is an example from Pereltsvaig (2004, p. 331):

(8) Sherlock Holmes: THE BUTLER did it!

In (8) *the butler* is contrastively focused, because he is different from the set of other potential suspects in the discourse. Also note that the focused noun, *the butler,* appears in all caps (true to the source material), showing emphasis on that noun. The use of emphasis in my criteria of topic and focus will be explored further in subsequent paragraphs. I will categorize general focus only, as further distinctions of focus are unnecessary to the analysis of my data.

Emphatic and Non-Emphatic Sentences

Another distinction to understand for this discussion of word order and discourse analysis is the difference between emphatic and non-emphatic sentences. King combines a couple of different studies on the subject, which will be briefly summarized here (Yokoyama, 1986). Non-emotive sentences are found particularly in writing and academic discourse more than in other registers. In a non-emotive sentence the topicalized constituents appear before the verb while focused constituents appear sentence-finally, such as an SVO sentence without intensifiers. However, a sentence is certainly emotive if the focus is anywhere else in the sentence. To clarify, in emotive sentences the focused constituent is likely found in preverbal position preceded by topicalized material. This understanding gives us another criterion for categorizing topic and focus in the data from the RNC.

In this chapter, I have surveyed the literature on Russian word order, discourse analysis, empirical methods and corpus studies, the Russian National Corpus, and notions of topic and focus. This information lays the groundwork for this thesis, which is discussed in greater detail in the next chapter.

Chapter 3: Corpus Data Sampling and Discourse Analysis Procedures

In the previous chapter, I discussed Russian word order and its possible constraints. Although Russian is sometimes said to have free word order, the literature proposes various constraints, without having a large set of data to validate the claims. This chapter will specifically address the method by which my research questions will be answered. I will discuss The Russian National Corpus and its architecture in greater detail, as well as the concepts of topic and focus and their role in the diagnostic criteria of my research.

Criteria for Categorizing Topic and Focus

In chapter 2, I surveyed a helpful dissertation, which provided useful insight into which criteria are important for this research in categorizing topic and focus (King, 1993). In this section of chapter 3, I will review the key points of those criteria, as well as the modifications I made to better suit the data I gathered from the Russian National Corpus.

A topicalized item is presupposed or old information. Topics tend to be definite and are often pronominal, and are always preverbal. King asserts that multiple topics and even no topic is possible in a Russian sentence. However, I found that in the types of sentences that King identifies as having two topics, one of the preverbal nouns was always mentioned in the preceding sentence, which leaves that last criterion unused. Verbs are discourse neutral, and were not included in my analysis of topic and focus.

Conversely, a focused item is considered new to the discourse and not presupposed. The focused item is most often found in the final position of the sentence. Additionally, King points out an exception to the final position criterion: in emotive sentences, the order is likely topicalized material first, focused material second, and verb last. This is not problematic, as we've seen that items which are neither the topic nor the verb can be identified as the focused

material. Lastly, the particle *zhe* in Russian is an intensifier, effectively marking the focus of a sentence.

These criteria from King, with a slight modification, allowed me to effectively analyze the data from the Russian National Corpus and categorize each sentence for the later statistical tests.

Architecture of the Russian National Corpus

As mentioned earlier, the Russian National Corpus (RNC) contains 149,357,020 word tokens taken from spoken, fiction, and written media (including academic and non-academic texts) of Russian from the mid-18th century to the present. Every text in the corpus is metatagged for information about the author, source, subject matter, etc., as well as morphologically tagged by computer. Additionally, each search result is displayed with several sentences of preceding and following context, which is very useful to discourse analysis. However, as discussed in previous chapters of this thesis, morphosyntactic annotation for Russian is difficult. Despite its problems, the RNC remains the best tool for answering the questions of this thesis.

Russian can exhibit a high amount of morphosyntactic ambiguity, which makes building a corpus of Russian very difficult. One word can have as many as 40 different forms, and there are many word forms that have one part of speech but an identical form can be found with another part of speech derived from a completely different root. This leads to high amounts of ambiguity and "noise" in the search results. Therefore the creators of the RNC devised a subcorpus called the Disambiguated Corpus. The Disambiguated Corpus contains 5.8 million tokens as of November 2007, and all of them are disambiguated by the creators. Each sentence in the Disambiguated Corpus (also called the Russian Dependency Treebank) uses dependency trees as the notation formalism. All morphological and syntactic ambiguity has been resolved

according to guidelines developed in the Laboratory for Computational Linguistics, Institute for Information Transmission, and Russian Academy of Sciences. Perhaps most importantly, because of its dependency trees, a query for "transitivity" is only available within the Disambiguated Corpus. This query function is vital to the methodology of this thesis research. The Disambiguated Corpus, although smaller than ideal, is the best option for answering the questions of Russian word order that this thesis will address.

It is important, also, to note the composition of data in the RNC. The researchers state on the homepage of the English version of the RNC website[10] that the distribution of texts is representative of the language usage of the time at which the corpus was created. In the RNC fiction texts total 39.7%, spoken is 3.9%, academic is 12.8%, and non-academic is 43.6%. The total composition of non-fiction texts is 56%, but for this thesis research I want to see the difference in register between general non-academic texts (which I refer to as "non-academic" throughout this thesis) and academic texts, so I categorize them separately. The Disambiguated Corpus is never explicitly specified as having a different composition of data from the main corpus, so I will assume that the data composition is the same. Although the balance of registers is not modeled after other large corpora, it may be a better representation of the language usage. Because my research is designed to get an approximation of the language usage, this makes the RNC the best corpus to use for this thesis.

[10] http://ruscorpora.ru/en/corpora-intro.html

Data Gathering and Analysis

Procedure

The procedure for gathering the data from the RNC is outlined in this section. On the main page of the English website, I chose 'search the corpus' from the options on the left side of the screen. The next page to appear shows the option to 'customize subcorpus' at the top right. After selecting 'customize subcorpus', the next page loads to show a list of different subcorpora. 'Disambiguated corpus' is the first option. Once I checked 'disambiguated corpus' I clicked the 'next' button at the bottom of the page. On the following page, I clicked the 'save subcorpus and search' button. The subsequent page shows the search query interface, with options for many different kinds of searches (see Figure 1).

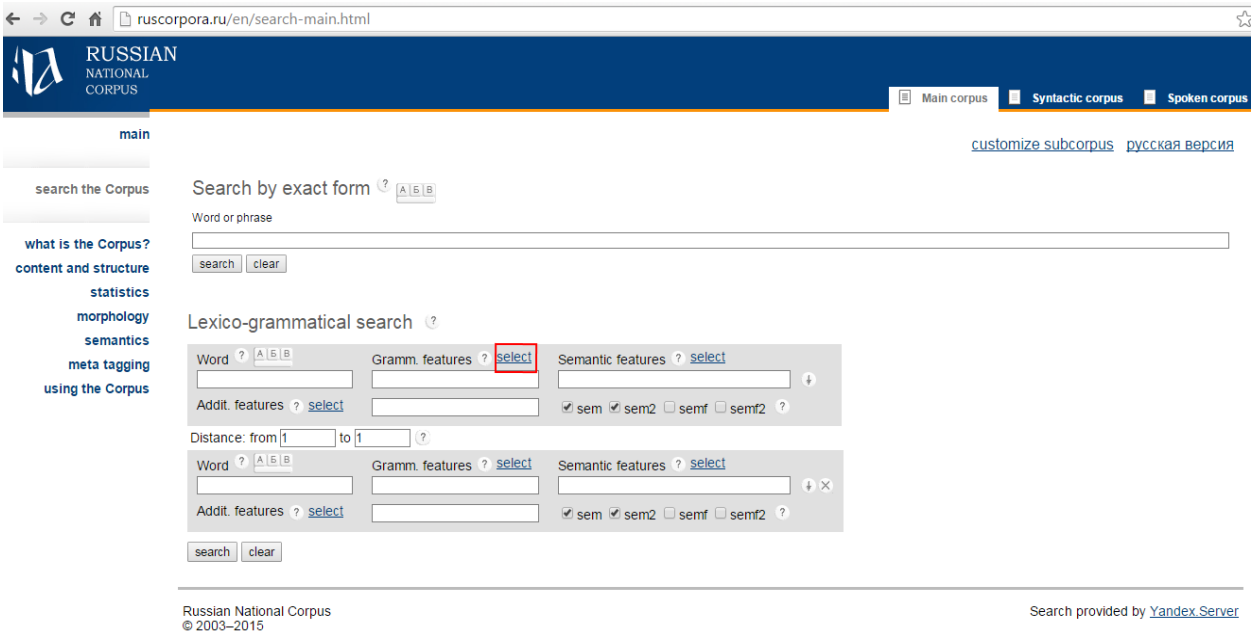

Figure 1. Screenshot of RNC Search Interface with Grammatical Features Query Highlighted

The search page of the RNC looks like this, but I edited the screenshot with the red box to show the select button for the grammatical features query. Once I clicked the grammatical features

'select' option, a popup window appeared, showing copious options. I checked the 'transitivity' box and clicked 'ok' at the bottom of the window (see Figure 2). Figure 2 shows the 'transitive' option marked under the grammatical features window. Again, the screenshot was edited with the red box to clearly show where the option is on the screen. The window closed, returning back to the search interface. After checking the appropriate box, I clicked 'search', which loaded the first of hundreds of pages of results. Next to each source was a link 'All examples' and a number next to it, showing how many transitive sentences were found in that source document. Once I clicked on that link, I could look through each individual instance within several sentences of preceding and following context[11]. The user interface of the site was somewhat difficult to understand, but once I learned the interface it allowed for this search procedure.

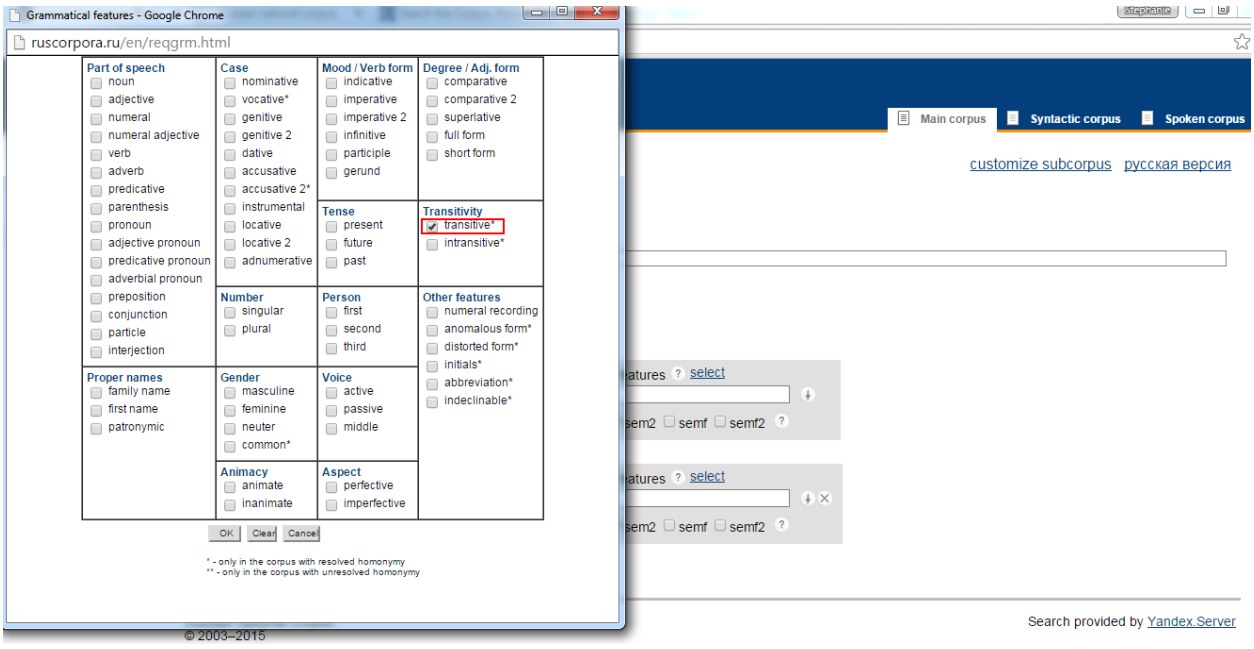

Figure 2. Grammatical Features Popup Window with Transitive Feature Selected and Highlighted

[11] Note: There was an option to download the entire batch of search results into XML format, but that did not allow for seeing each instance with the context of at least the preceding sentence, which context is vital to the present study. Without downloading the search results, it was necessary to reenter all the search parameters in the website every time I analyzed the data. Fortunately, the exact same results in the exact same order were returned every time I used the corpus in a six-month period, so the search procedure remained valid.

Search Criteria

Once the search results were available to view in the browser window, I proceeded sequentially down the list as it was returned by the search until I had 500 sentences that worked for my research questions. I either included or excluded the sentences into the analysis based on several additional criteria. I included only instances of transitive sentences which had at least one preceding sentence in the source material[12]. This criterion applied the Principle of Local Interpretation, which says that only the most immediate context is needed to interpret the sentence (Blekher, 1995). Blekher reasons that this means only the preceding sentence is needed to interpret the meaning, and therefore also to determine the topic and focus of each sentence. Additionally, some of the instances included in the search results contained ditransitive verbs, which were not suitable for the analysis—each sentence required one each of subject, object, and verb. Double verb constructions (such as "It allows us to see…") were also excluded for the same reason. Furthermore the search results included transitive verbs from both main clauses and embedded clauses, but only main clause sentences were included in the analysis. The results also returned non-verbs, such as participles, which were not suitable for the analysis and were thus excluded. Only indicative verbs were included, as other moods often exhibit radically different behavior. Also all 'Subjectless Dative Constructions' were excluded, as were other sentences with covert subjects or covert objects: the present study deals only with overt thematic roles. An example of a Subjectless Dative construction is the following:

(9) *Mamye* *nado* *gotovit* *uzhin*
Mother.DAT necessary cook.3sg dinner
'It is necessary that mother cooks dinner.'

I also excluded any verbs with the –sya suffix, such as the simple reflexive sentence:

[12] Some sentences that I excluded were the first sentence in the source document, and thus had no preceding context.

(10) *Pasha* *umyvaet- sya*
Pasha.NOM wash.3sg-REFL
'Pasha is washing up'

The polysemantic –sya corresponds to reflexive, passive, reciprocal, and other meanings. This suffix and its nuanced uses were beyond the scope of this thesis.

Lastly, any punctuation or spelling that rendered a sentence unreadable caused that instance to be excluded as well. These criteria ensured that the instances could be encoded into an Excel spreadsheet for categorizing and statistical analysis.

Topic and Focus Analysis

After I found the 500 sentences that met the criteria for inclusion, I analyzed each one for the positions of topic and focus. In this section I will demonstrate how I determined the topic and focus of each sentence in the data sample. The following examples from my data sample will show the step-by-step process. The first example I will use is the following sentence:

(11) *StorEdge 3511* *podderzhivaet raznorodnye* *operatsionnye*
StorEdge 3511 support-3SG heterogenous.ADJ.PL operation.ADJ.PL
platformy *i* *razlichnye* *sredstva* *clasterizatsii serverov*
platform.PL and variety.ADJ.PL resource.PL.GEN clusters.PL server.PL.GEN
'StorEdge 3511 supports heterogenous operating platforms and a variety of clustering service resources'

The previous sentence to (11) in the source text was the following:

(12) *Eto* *oznachaet,* *chto systema* *mozhet ispol'zovat-sya* *v* *usloviyakh*
This mean.3SG that system can use.INF-REFL in condition.PL.PREP
samykh *zhëstkikh* *vneshnikh* *vozdeystviy*
most.PL.PREP tough.PL.PREP external.PL.PREP influence.PL
'This means that the system can be used in the most demanding conditions of external influences'

The subject in sentence (11) is 'StorEdge 3511', which is coreferential with 'the system' in sentence (12). By King's criteria, this means that the subject of sentence (11) is the topic. The object in sentence (11) is 'heterogeneous operating platforms and a variety of clustering service

resources', which is not coreferential with anything in the preceding sentence, (12). This means that the object is the focus of the sentence. Sentence (11) is SVO, and displays the subject topic before the object focus. Both the subject and the object in this sentence were full noun phrases, which was not always the case in this data sample.

The next non-academic example sentence (13) contains a pronoun subject instead of a full noun phrase.

(13) *Oni predlagayut kak servernye materinskie platy*
3PL offer.3PL such server.PL mother.PL board.PL
'They offer such server motherboards'

The pronoun 'they' is coreferential with the appositive 'Tuan and Supermicro' in the previous sentence, (14).

(14) *V etoy nishe predstavleny, po suti, dve osnovnye*
In this niche represent.PST.PL.PTCP.PASS in essence.PL two main.PL
Kompanii Tuan i Supermicro
company.PL Tuan and Supermicro
'In this niche are represented, as a matter of fact, two main companies—Tuan and Supermicro'

By King's criteria, topics are often pronominal and preverbal. Thus the topic of (13) is the preverbal subject, 'they'. The object in (13) is new to the discourse (not coreferential with anything in the previous sentence) and is in final position, so by the criteria, the object performs the function of focus in this sentence. Again, sentence (13) is an SVO sentence with the subject topic before the object focus, which represents the majority of the sentences in the sample of 500 sentences from the corpus.

Spreadsheet and Categorization

The method of categorizing and encoding the data into the spreadsheet will be detailed in the following section. I included columns in the spreadsheet for three groups of information:

discourse topic and focus, the six possible word orders, and the four registers. After utilizing the criteria for analysis listed in previous sections of this paper, I entered the information into the spreadsheet. I encoded each piece of information about a sentence as 0 if the sentence didn't display that information, or 1 if it did. The spreadsheet included 14 columns total for the information encoding and 500 rows: one row for each sentence that was encoded. To clarify, each included sentence had one topic and one focus per sentence, therefore the subject of the sentence could be either the topic or the focus, and the object of the sentence could be either the topic or the focus. If a sentence was SVO, it was marked 1 under the SVO column and the other word orders are marked 0, and so on for the other word orders. Lastly, if a sentence was from the academic register, I marked the Academic column with 1 all other registers were marked 0, and so on. Although encoding the data in the spreadsheet this way was initially somewhat difficult, it was necessary for the statistical analyses that I subsequently performed. It allows the mutually exclusive data values from all categories to be independently calculated.

Statistical Test

After totaling the frequencies of each group of data, I needed to determine whether the results were a product of chance or whether they are statistically significant. I used a chi-square test to accomplish this, as similarly used in Smith (2013). This test is performed when the observed frequencies of different categories are compared to a set of expected values and it returns values which indicate whether or not the differences were a product of chance. The expected values are calculated to reflect the null hypothesis, and the test procedure calculates the aggregate difference between the expected values and the observed values in the data set.

The goals of my thesis were to see whether the different word orders would occur in the same proportions. Additionally I wanted to see whether the word order variations were

influenced by the positions of discourse topic and focus, and whether the variations were influenced by register. If the chi-square test showed values that were statistically significant, it would indicate that the observed frequencies of the different word orders likely were influenced by discourse patterns and register. If the chi-square test did not yield statistically significant values, I would not be able to say whether the data could have been produced by chance. The totals for each category and the corresponding statistical tests will be discussed in Chapter 4.

Chapter 4: Results

In this chapter, I present the results of the data analysis and statistical tests. The questions of this thesis are:

1. If all six possible word orders are possible, how frequently would each form occur in the given data set from the corpus?
2. Does the position of discourse topic and focus influence the variations in word order frequency?
3. Does register (spoken, fiction, academic, or non-academic) influence the variations in word order frequency?

I will address each question with the resulting data in the paragraphs that follow. I also show example sentences from the corpus data sample. The complete list of all 500 sentences and my annotations used in the analysis is available via download.[13]

Word Order Frequencies

Firstly, if all six word orders are possible, what are the frequencies of each word order in the corpus data sample? The data sample gathered from the corpus totaled 500 transitive sentences. Out of those 500 sentences there were 448 SVO; 12 OVS; 22 SOV; 8 VSO; 1 VOS; 9 OSV. These values, along with their percentages of totals, are shown in Table 1.

As predicted in the literature, all of the possible word orders were found in the data sample. SVO sentences made up the vast majority at 89.6% of the total sentences from the corpus sample. SOV sentences were the next highest in frequency with 4.4% and OVS was third with 2.4%. Interestingly, these results are similar to those previously discussed (Kallestinova,

[13] http://linguistics.byu.edu/thesisdata/BillingsRussianData.xlsx

2007). Kallestinova found from grammaticality judgments and elicited sentences from 237 native speakers that the orders SVO, OVS, and SOV were preferred. The other word orders were not produced by the speakers, but regarded as acceptable.

Table 1. Attested Word Orders from the Corpus Sample in Numbers and Percentages

Word Orders	*# of sentences*	*% of sentences*
SVO	448	89.6%
OVS	12	2.4%
SOV	22	4.4%
VSO	8	1.6%
VOS	1	0.2%
OSV	9	1.8%

In this thesis data SVO, SOV, and OVS word orders account for 96.4% of the sentences from the sample. It seems that these three word orders are indeed greatly preferred by speakers. Although VSO, VOS, and OSV together comprise 3.6% of the data sample, each order was attested in the data at least one time.

Statistical analysis of frequency data

I performed a chi-square test to determine the likelihood that the observed frequencies in the data were a product of chance. To perform the test, I calculated expected values for each word order based on my null hypothesis, which is that all word orders are equally likely. This meant that the expected values would be evenly distributed across the six categories. The test compares the expected values to my observed values from the data set, and finds the aggregate

difference between expected and observed. The results of the test indicate that the difference is statistically significant at the .001 level (χ^2=1917.812, *df*=5, N=500, *p*< .001).

Based on the results of the chi-square test, I can make reliable inferences about general word order patterns in Russian. 89.6% of the sentences in the data sample were SVO sentences. Such a large statistically significant majority provides evidence for a predominant Russian word order; the data shows that SVO is the preferred order. However, due to the occurrence of every possible word order in the data, it is still fair to say that any word order is possible in Russian, although some are less frequent.

Position of Discourse Topic and Focus

The second research question of this thesis asked whether the position of discourse topic and focus influenced the variations in word order frequency. I will present the findings of my analysis and the statistical test.

Topic and focus data

I had originally designed the columns to allow for the encoding of sentences that had only a topic or only a focus, according to King's criteria. This meant that I had four discourse columns in the spreadsheet: Subject Topic, Subject Focus, Object Topic, and Object Focus. However, in my analysis I found that each sentence had exactly one topic and one focus, so I modified King's criteria for the analysis. When the subject was the sentence topic, the object was always the focus. Conversely, when the subject was the sentence focus, the object was always the topic. These corresponding data points made the four columns redundant in the statistical tests, so I collapsed them into two discourse categories: Subject Topic and Subject Focus. After encoding all the data points into the spreadsheet, I totaled the different columns and rows to

show how often each word order occurred with the two discourse patterns. These categories are labeled S topic and S focus in Table 2.

In the data sample, word orders in which the subject precedes the object (SVO, SOV, VSO) are more likely to have a subject topic. Similarly, word orders in which the subject follows the object (OVS, VOS, OSV) are more likely to have a subject focus.

Table 2. Totals of Subject Topics and Subject Foci across Word Orders

Discourse	*SVO*	*OVS*	*SOV*	*VSO*	*VOS*	*OSV*	*Total*
S focus	11	10	1	1	1	7	31
S topic	437	2	21	7	0	2	469

In the entire data sample overall, the topic precedes the focus in 465/500 or 93% of sentences. To further illustrate this point, Table 3 displays the percentages of subject topics and foci within each word order.

Table 3. Percentages of Subject Topics and Subject Foci within Each Word Order

	SVO	*OVS*	*SOV*	*VSO*	*VOS*	*OSV*	*Total*
S focus	2.5%	4.6%	12.5%	83%	100%	77%	7%
S topic	97.5%	95.4%	87.5%	17%	0%	23%	93%

This shows that in the data sample, topics often occur first (and focused constituents often occur last) in a sentence. This makes sense, as King and many other researchers assert that the topic is usually the first constituent in the sentence and the focus is usually the last.

The data reviewed in this section indicate that, indeed, the positions of discourse topic and focus seem to influence word order variations; the topic is most often first in the sentence and the focus is most often last. But can I say that these patterns did not emerge in the data as a

product of chance? The statistical analysis will tell whether the results are statistically significant.

Statistical analysis of discourse pattern data

I implemented a chi-square test to determine whether the results are statistically significant. I calculated the expected values based on the null hypothesis, which is that discourse patterns do not influence word order variations. Thus the expected values would be evenly distributed across the two discourse patterns. The chi-square test showed that in this case the differences between the categories is statistically significant at the .001 level (χ^2=4101.128, *df*=5, N=500, p< .001). This indicates that in the data sample, the discourse patterns did influence the word order variations.

Register

The third research question of this thesis asked whether register influenced the variations in word order frequency. I will present the findings of my analysis and the statistical test.

Register data

The design of the spreadsheet included four columns, one each for spoken, fiction, academic and non-academic. In my sample of 500 sentences from the corpus I did not find any sentences from either the spoken or fiction registers; therefore spoken and fiction are not shown in the table below but they were included in the statistical test. The distribution of texts between the different registers is shown in Table 4.

Table 4 . Word Order Variation Totals between Registers

	SVO	*OVS*	*SOV*	*VSO*	*VOS*	*OSV*	*Total*
Academic	119	2	2	4	1	0	128
Non-Academic	329	10	20	4	0	9	372
Total	448	12	22	8	1	9	500

At a glance, it is obvious that non-academic texts outnumber academic texts in the sample: non-academic texts comprise 74.4% of the total sample of 500 sentences. Also note that non-academic texts occur more often for most of the word orders: the occurrences of each register are shown within the word orders in Table 5.

Table 5. Percentages of Register Occurrences within Each Word Order

	SVO	*OVS*	*SOV*	*VSO*	*VOS*	*OSV*
Academic	26.56%	16.67%	9.09%	50%	100%	0%
Non-Academic	73.44%	83.33%	90.91%	50%	0%	100%

These percentages imply that register may have an influence on the variations of the word orders. The percentages may also be caused by other factors, however, which will be detailed in the discussion chapter. From the data alone, it seems that register may influence the variations of word order, but I cannot yet be certain. The statistical analysis will give an estimate of the likelihood that the results were obtained by chance.

Statistical analysis of register data

I performed a chi-square test to determine whether the results of the register data analysis were statistically significant. I calculated the expected values based on the null hypothesis that register does not influence word order variations. The expected proportions of each register reflected the proportions reported for the corpus. The chi-square test showed that in this case the differences between the categories is statistically significant at the .001 level (χ^2=3825.3, *df*=15, N=500, $p < .001$). This indicates that register did influence the word order variations in the sample.

Example Sentences from Corpus Sample

I will now show examples of all word orders from my corpus data sample, as well as examples from both registers. For SVO sentences, please see Sentences (11) and (13) in the methodology chapter. An academic OVS sentence from the sample is (15):

(15) *Sevodnya nas posetil President-Ø Rossii*
Today 1PL.GEN visit.PST.MASC President-NOM.MASC Russia.GEN
'Today the President of Russia visited us'

An example of an academic SOV sentence from the sample is (16):

(16) *My ochen' rad-y Vas videt'*
1pl.NOM very glad-PL 2pl.GEN see.INF
'We are very glad to see you (all)'

An academic VSO sentence from the sample is (17):

(17) *Poblagodarila Neschastnaya Roza Prekrasnuyu*
Thank.PST.F Unhappy.ADJ.F Rose.NOM.F Beautiful.ACC.F
Mariannu za beluyu yakhtu
Marianne.ACC.F for white.ADJ.ACC yacht.ADJ.ACC
'Unhappy Rose thanks Lovely Marianne for the white yacht'

An academic VOS sentence from the sample is (18):

(18) *Raskryl Sherlock Holmes eto slozhnoye delo*
Discover.PST.M Sherlock Holmes this.N difficult.N matter.N
'Sherlock Holmes discovered this difficult matter'

A non-academic OSV sentence from the sample is (19):

(19) *Druguyu polovinu galakticheskogo goda Aristotle provodil*
Other.ACC half.ACC galactic.GEN year Aristotle spend.PST
'The other half of the galactic year Aristotle spent'

These sentences demonstrate that all six word orders occurred in the data sample from the corpus, and that both academic and non-academic registers also occurred in the sample.

In the methodology chapter, I discussed the process of determining topic and focus in two example sentences, (11) and (13). Both of those SVO sentences displayed the topic before the focus. In this paragraph I will show an SVO example from the corpus sample that displayed the pattern of focus before topic. This example is (20):

(20) *Zhilishchnyy vopros ne osobenno zabotit starsheklassnikov*
Housing question not especially worry.3SG upperclassmen.PL.GEN
'The housing question does not especially worry upperclassmen'

The sentence in the sample previous to (20) provides the context for determining the topic and focus of (20). This contextual sentence is (21):

(21) *Pri etom 57,6% starsheklassnikov schitayut nevozmozhnym*
In this.PREP 57.6% upperclassmen.PL.GEN find.3PL impossible
zarabotat' den'gi v svoëm gorodye
earn.INF money in self.PREP.M city.PREP
'In this case 57.6% of upperclassmen find it impossible to earn money in their own city'

The noun phrase (NP) 'upperclassmen' is mentioned in (21), which is coreferential with the NP 'upperclassmen' in (20). This makes 'upperclassmen' in (20) the topic, although it is the object and in final position. The focus in (20) is the NP 'housing question' because it is not coreferential with anything in its preceding sentence and is therefore new to the discourse. In addition, the example sentences have shown the use of full NPs as well as the use of pronouns.

This illustrates the pattern of both full NPs and pronouns in the sentences throughout the entire corpus sample.

Chapter 5: Discussion

The results chapter illustrated that the data showed statistically significant patterns. For example, in the sample of 500 sentences, the vast majority were SVO sentences. Also congruent with previous research on the subject, Russian discourse patterns follow the traditional discourse patterns (topic before focus in a majority of occurrences) and that discourse patterns influence word order variations. The analysis of the divisions between the different registers yielded somewhat puzzling data, although the data was statistically significant. In this chapter I will review some of the potential reasons for the patterns in the data, how well my research questions were answered, and possible interpretations of the results. Finally, I will discuss the usefulness of the RNC as a research tool for this thesis.

Word Order Frequencies

The results of the word order frequency analysis showed that the majority (89.6%) of sentences in the sample of 500 sentences from the corpus were SVO. The analysis also showed that SOV (4.4%) and OVS (2.4%) were next in frequency, and that every word order occurred at least one time in the sample.

I observed some things about the data that may explain the patterns. The 500 transitive sentences obtained from the corpus at large may not be a true representation of the language for a number of reasons. Firstly, the corpus has a much different balance of registers than other large corpora, although every corpus is balanced differently. The creators of the RNC reason that the distribution of texts represents the language usage of the time at which the corpus was created. The composition of texts in the both the main corpus and disambiguated corpus is as follows: fiction texts total 39.7%, spoken 3.9%, 12.8% academic, and 43.6% non-academic. By extension this means that the raw probability of finding a sentence from the spoken register in a sample of

500 sentences is quite low. This may mean that the data from my sample is somewhat skewed for register, but it may be representative of the actual usage.

Secondly, the sentences gathered are not representative of the corpus because of the myriad criteria for analysis. Namely, only active transitive sentences with overt thematic roles were included[14]. Active transitive sentences with overt thematic roles tend to exclude some instances of informal speech and discourse situations in which marked word orders may be more likely. For instance, Bresnan (1994) claims that locative inversion in Russian only occurs in intransitive sentences or sentences with passivized transitives. Interestingly, locative inversion is frequent at the beginning of a text sample or a paragraph because it often signals the start of discourse with information about when and where. I had to exclude samples that had no preceding context, and this may be another reason that my analysis did not include instances of locative inversion. The subsequent exclusion of locative inversion and makes for potentially fewer instances of marked word orders in the data analysis. Additionally, "subjectless dative" constructions and sentence fragments were also excluded for lack of overt thematic roles. These constructions occur frequently in informal speech, and marked word orders are more likely in less formal contexts. However, adding these excluded constructions back into the analysis was beyond the scope of this thesis.

Thirdly, I did not find any sentences in the 500 from spoken or fiction registers that met the criteria. Based on what I know about the overall composition of the corpus, this is surprising. Similarly to what I have discussed so far, having data that is not balanced for register is somewhat disappointing for the analysis; my claims about the influence that register has on word

[14] Russian passives are incredibly nuanced both morphosyntactically and semantically. Their inclusion would have inordinately complicated the analysis and results. The inclusion of covert thematic roles was also beyond the scope of this thesis.

order variations can only reference two of the four registers in question. Data from the spoken and fiction registers may have even shown more occurrences of the more marginal word orders.

Fourth, my research methodology may have skewed the results. The methodology that I designed relied on the part-of-speech tagging that returned only transitive results from the search function. Without using the part-of-speech tagging, I had the option of using the top ten most frequent transitive verbs in Russian as a search parameter. I chose not to use that option because I understood it to be less representative than searching for all transitive verbs in the corpus. Alternatively I could have searched through the raw data in the corpus without any search parameters but the amount of time that it would have taken to accomplish the research was beyond the scope of this thesis. Unfortunately, the RNC website did not have information about how the search function sampled the different registers from the main corpus and in what order (if any) those registers were displayed in the search results. This lack of information combined with how I gathered my sentences (I analyzed the first 500 sentences from the search results that fit my criteria) may also have contributed to the lack of spoken or fiction texts in my data sample. Despite the imperfections of the data sample, it still returned statistically significant results that proved useful to my thesis.

Although the data that I gathered may be limited by the different balance of registers in the RNC, the large majority of SVO sentences in the sample is a strong indicator that SVO is the preferred word order. In spite of the common idea that Russian is a language with free word order, in my review of the literature I found that many researchers agree that some word orders are preferred over others. Some researchers even go as far as positing that SVO should be considered the basic word order of the language. In my sample not only did I find a majority of SVO sentences, I also found at least one occurrence of each word order. I can reasonably infer

from this data that although any word order is possible in Russian, the basic word order is SVO. These inferences show that the word order of Russian is somewhat fixed.

Discourse Patterns Discussion

The results of the data analysis of discourse patterns showed that discourse patterns influence word order variations in the sample. Additionally I found that the 93% of the sentences exhibit a subject topic, and 95.6% of the sentences exhibit word orders in which the subject precedes the object.

This shows that the majority of the sentences in the sample display the topic before focus. These findings imply that discourse patterns influence word order variations. As stated in the results chapter, these findings are congruent with the previous assertions about Russian discourse patterns that I surveyed in the literature review. Based on the statistical significance of the results, I can infer from the data that discourse patterns influenced word order variations in my data sample.

Register Discussion

The analysis of the data from the corpus showed that the majority of samples were from the non-academic register. Specifically, 372/500 or 74.4% of the samples were from non-academic sources while only 25.6% were from academic sources. Also noteworthy was the lack of any samples from either the spoken or fiction registers. The three most common word orders (SVO, SOV, and OVS) showed a clear majority of occurrences from non-academic sources. The statistical analysis showed that the data are significant, but there are some noteworthy things about the register data that still warrant discussion.

Firstly, the patterns in the register data were unexpected. In the corpus, spoken texts total 3.9%. The expected values that I calculated for the chi square test were based on the proportions of registers in the corpus; for example, the expected value for the spoken register was 19.5 occurrences. I was surprised that the sample did not yield any fiction occurrences, as fiction comprises 39.7% of the corpus. The higher proportion of fiction texts is because Russian language experts often assert that literary Russian is the standard. So much so, in fact, that in two earlier Russian corpora as well as the RNC, the researchers included a higher proportion of literary texts than is found in either the Brown Corpus or the BNC. In speaking of this difference Sharoff (2006, p. 170), one of the creators of the RNC says "This reflects the difference in the cultural status of the language of imaginative writing in British and Russian cultures: in Russian the literary language is treated as the authoritative source, which effectively defines the language used by native speakers." Given the higher proportion of fiction/literary texts and the fact that they are considered the standard for language use, it remains unclear why I was unable to find fiction occurrences in my sample of 500 sentences from the corpus. I could have sampled 20–30 sentences from the spoken subcorpus of the RNC, but it was beyond the scope of this work. In the sample I did include many instances of both pronouns and full NPs, therefore the weight of an NP was not shown to be a factor in the lack of spoken or fiction registers in my sample.

Secondly, the majority of sentences in the sample were SVO from the non-academic register. This majority leads to very small numbers in some of the other categories, especially VSO and VOS. It is possible that a larger sample size may have changed the proportions of occurrences between the word orders.

The chi-square test showed that the results of the register analysis were significant. It would be more illuminating to have data from all four registers, but the test shows that register does influence word order variations.

State of Research Questions

Did the results of the data analysis sufficiently answer the research questions of this thesis? I will deal with the three questions individually.

Firstly, if all six word orders are possible, how frequently would each form occur in the given data set from the corpus? I find that this question was satisfactorily answered. I successfully gathered the amount of data that I wanted, analyzed the word orders, and totaled each order's frequency. The result was a clear division of sentences between all six possible word orders, with a majority of SVO sentences and at least one occurrence of every word order. I can also reasonably claim from the chi-square test that SVO seems to be the basic word order of Russian. I am satisfied with how that question was answered within the corpus.

Secondly, does the position of discourse topic and focus influence the variations in word order frequency? From the data that I gathered, it was clear that discourse patterns may influence the word order variations, and the chi-square test confirmed that discourse patterns certainly influence the variations. This research question was satisfactorily answered within the corpus.

Thirdly, does register influence the variations in word order frequency? The data by itself was unclear, but the chi-square test confirmed that the two registers I found do influence word order varations. However, I would have preferred to see data from the spoken and fiction registers in order to have a better picture of the actual usage.

All three of my research questions were answered within the corpus. However, the puzzling results of my analysis of register may have revealed a flaw in the RNC. The implications of these factors for future research will be further discussed in the conclusions chapter.

Usefulness of the RNC

In this section I will discuss the Russian National Corpus as a research tool and more specifically, its usefulness to this thesis.

The RNC is a large corpus containing 149,357,020 word tokens, each with robust morphological annotations and metatagging. Each RNC search result is displayed with several sentences of preceding and following context. The RNC also includes the Disambiguated Corpus, containing 5.8 million tokens which were disambiguated by hand. The function to search for transitive sentences is only available within the Disambiguated Corpus. This thesis research dealt only with overt transitive sentences, and I needed to see the context of each sentence for discourse analysis purposes. In spite of its small size, my choice to use the Disambiguated Corpus was not shown to be a factor that limited the answers to my research questions.

The composition of the corpus may have created problems for my research. In the corpus, fiction texts total 39.7%, spoken is 3.9%, academic is 12.8%, and non-academic 43.6%. The 3.9% of spoken texts in the RNC resulted in no occurrences of spoken texts in my sample of 500 sentences, and mysteriously there were no fiction occurrences, which left my third research question without a lot of the anticipated data.

In addition, the search functions of the RNC may have skewed the data. When I analyzed that data for occurrences of different registers, I was unable to find any instances of fiction in my

sample of 500 sentences. This was surprising, considering that the percentage of fiction texts in the corpus is 39.7%, which is well above the proportion displayed in some other large corpora. This result was completely unexpected, and by all indications on the website for the RNC, should not have happened. In hindsight, it may have been possible to devise a different search method that would have gathered equal amount of data from each register of the corpus. The corpus contains a small subcorpus of spoken texts, as well a search function that allows you to narrow your fiction search results by 11 different genres. There is a similar function for non-fiction texts, in which you can narrow your results to 18 different sub-types, each of which is further divided into several more specific categories. Based on these search functions, I could have instead gathered in four different searches an equal amount of data from spoken, fiction, academic, and non-academic registers. However, my research design was intended to get a random sampling of the entire corpus in order to understand patterns in the language overall. A different methodology would have warranted a change of my research questions to suit it. Therefore the methodology that I originally devised remains the best to answer my research questions: instead the search results were not as representative of the corpus as I initially thought they would be.

The state of the RNC is less than ideal, but unfortunately, it is a reflection of the current state of Russian corpus linguistics. In review, the morphological complexity of Russian leads to high amounts of ambiguity between forms and makes it impossible for researchers to use the same taggers that have been developed for corpora of other languages, like English or Romance languages. The vastly different tagging method makes computer parsing difficult, leading to "noise" in the search results. This difficulty of computer parsing means that the RNC is currently the only large corpus that has a remotely balanced composition of texts as well as morphological

annotation or lemmatization. There are smaller corpora, such as the Uppsala Corpus: the balance of the 600 texts that comprise the total 1 million tokens is 50% fiction, 50% non-fiction, but it is not annotated nor lemmatized. Additionally, there are some internet corpora that are quite large (20 million or more tokens), but are limited to only internet texts and don't have as much morphological annotation as the RNC. Copyright laws prohibit use of extended context in corpora texts, which adds another degree of difficulty for researchers in gathering texts. Russian corpus linguistics is behind other languages in the field, so a researcher's options are limited.

My main research objective was to get a snapshot of the actual usage of the different word orders in Russian. This led me to devise secondary research questions to try and explain the possible results of the first question. My review of the literature pointed to discourse patterns and register being two likely factors in word order variations. Thus I chose to use the biggest and most balanced annotated corpus of Russian that I could find. Similarly, I designed the methodology to try and get a random sampling of the corpus, and hopefully an approximation of the language usage overall. My research questions were answered satisfactorily, but some of the results could not have been predicted. In spite of its imperfections, I chose to use the RNC because it was best suited among my other options for the research questions of this thesis.

Chapter 6: Conclusions

There are a variety of reasons to study Russian. Firstly, it is one of the most commonly spoken languages in the world, with an estimated 166 million speakers. Secondly, Russian is morphologically synthetic, meaning that one word is made up of multiple or several morphemes. Thirdly, Russian employs a rigid system of case marking with six cases of noun and adjective declensions that also convey person, number, and gender. The few exceptions to the case marking are mostly the indeclinable nouns, which are typically foreign borrowings. This system of case marking leads to freedom in the word order; all six word orders (SVO, OVS, SOV, VSO, VOS, and OSV) are said to be possible in Russian. However some researchers assert that certain word orders occur more than others do, and that other orders are even considered marginally acceptable by native speakers. This thesis aimed to perform introductory research to find out more about the frequencies of Russian word order in actual usage.

There are many studies from the last few decades that make claims about word order patterns in Russian and the possible reasons for the variations. Reasons like discourse patterns, register, constituent weight, and intransitivity constraints were among them. Unfortunately, many of the studies utilize small data sets, many of which are formal, meaning that the example sentences come from the researcher's mind and are more likely to be biased. This could not be helped; for decades after Chomsky's famous *Syntactic Structures* in 1957, corpus research has been fairly obscured and regarded as somewhat unreliable in linguistics. However, modern advances in computing have made corpus linguistics viable and increasingly popular as research tools. Access to larger and more reliable data sets from corpora is becoming widespread, and more papers are being published that show actual usage data obtained from corpora. The academic literature of Russian linguistics does not have many published papers that utilize

corpus data, and I was unable to find a paper that used corpus data to show actual word order usage patterns in Russian. Upon seeing this gap in the literature, I resolved to design and execute an introductory corpus study that would gather information about word order frequencies and investigate the possible effects of discourse patterns and register on those variations.

While designing the methodology, I found that the reason for the lack of corpus studies of Russian was due to the existence of only a few freely available Russian corpora. I decided that the Russian National Corpus suited the needs of my thesis better than the rest. I gathered 500 transitive sentences from the Disambiguated Corpus and analyzed them to see how many of each word order were in the sample, as well as document the discourse pattern and register shown in each sentence. My analysis showed that 89.6% of the 500 sentences were SVO, and that 93% of the 500 sentences displayed the "topic before focus" pattern that many other languages generally display. The statistical test showed that the observed word order variations are significant and that the discourse pattern data is statistically significant. Additionally my analysis found data from only two of four registers of Russian yet the statistical test showed that the data regarding those registers was significant.

I inferred from the results and the tests that SVO can be considered the basic word order of Russian. I also inferred that the discourse patterns and register influence the word order variations. Additionally I concluded that the limitations of the RNC are more serious than I originally thought them to be; a different sampling methodology may have yielded more complete data, although the research questions would have to be slightly different as well.

Future Research

From this thesis research I learned many things about discourse analysis, word order variations, corpus linguistics, statistical analysis, and many other things that lend themselves

well to future research. I will highlight some of the potential future studies that I think would be valuable.

First, I found in my analysis of the corpus data that there were some sentences that had either a covert subject or covert object. I excluded these sentences from my analysis because they did not answer my research questions about word order. However, in a future study I would be interested to see the frequencies of covert subjects and covert objects, especially when those frequencies are compared against discourse patterns and register.

Second, I excluded many sentences that were "subjectless dative" constructions. This construction is a controversial topic among Russian language experts, and a corpus study could lend more insights into the semantic interpretations of this puzzling construction.

Third, one of the factors that may influence word order variations that I did not deal with in my study was the proposed Intransitivity Constraint of Locative Inversion (Bresnan, 1994). The data that I obtained from the corpus were all transitive sentences, so according to this constraint, I may have excluded from my analysis occurrences of other word orders besides SVO. A different search methodology to allow the inclusion of intransitives and passivized transitives may be illuminating.

Fourth, sampling differently from the corpus should be a focus for future work. For example, sampling 125 sentences from each register of the corpus would yield more balanced results, and would likely solve the mysterious problem of register data that I encountered in my methodology. It would eliminate the possibility of the search function returning search results from the main corpus that are skewed for register. Likewise, it would solve the potential problem of the balance of registers in the main corpus. Alternately, a study that compares the different

corpora of Russian would be useful: for instance, the design could include sampling from only the internet genre of the RNC and comparing the results to the same search from an internet corpus of Russian.

Fifth, my corpus sample of 500 sentences included both full NPs and pronouns. Any further analysis of constituent weight was beyond the scope of this thesis. However, a future study that analyzes constituent weight (heavy vs. light or full NPs vs. pronouns) as a factor in word order frequencies would be interesting. It would be informative to see if the constituent weight in Russian affects the frequencies of certain registers in a corpus sample, e.g. if pronouns are more common in the spoken register.

Sixth, this thesis unintentionally highlighted some of the weaknesses of the RNC. Improved search functions that allow for even more robust searches would greatly help future researchers. Both the English and Russian interfaces of the corpus are not very user-friendly; I was well versed in how to use COCA, other corpora, and part-of-speech tagging before I performed this study yet so much of the design and interface of the RNC was opaque to me at first. It took many hours of poring over both the English and Russian versions of the website and doing countless sample searches before I understood enough of the search functions and what they meant. Someone with less background knowledge than I had might find this interface unintelligible. Finally, adjusting the balance of the registers in the corpus may result in more complete data in the future.

This thesis illuminated much of the current state of Russian corpora in general and the RNC specifically. This introductory work was a good start, but there is still much to improve and learn in this field.

References

Alekseyenko, N. V. (2013). *A Corpus-Based Study of Theme and Thematic Progression in English and Russian Non-Translated Texts and in Russian Translated Texts.* (Doctoral dissertation). Kent State University, Kent, OH. Retrieved from ProQuest Dissertations and Theses. (AAI3617718)

Apresjan, V. (2013). Corpus Methods in pragmatics: The case of English and Russian Emotions. *Intercultural Pragmatics, 10*(4), 533-569.

Blekher, M. (1995). Word Order and Discourse Management in Russian. *University of Alberta Papers in Experimental and Theoretical Linguistics, 3*, 1-14.

Brennan, S., Friedman, M., & Pollard, C. (1987). A Centering Approach to Pronouns. *Proceedings of the 25th Annual Meeting of the ACL*, (pp. 155-162). Standford, CA.

Bresnan, J. (1994). Locative Inversion and the Architecture of Universal Grammar. *Language, 70*(1), 73-131.

Chomsky, N. (1957). *Syntactic Structures.* The Hague: Mouton.

Davies, M. (2008). New Directions in Spanish and Portuguese Corpus Linguistics. *Studies in Hispanic and Lusophone Linguistics, 1*(1), 149-186.

Dyakonova, M. (2004). Information Structure Development: Evidence from the Acquisition of Word Order in Russian and English. *Nordlyd, 32*(1), 88-109.

Dyakonova, M. (2009). *A Phase-Based Approach to Russian Free Word Order.* (Doctoral dissertation). University of Amsterdam, Amsterdam, Netherlands. Retrieved from http://www.lotpublications.nl/Documents/230_fulltext.pdf

Fiedosjuk, M. (2010). "As Industrious as a German and as Light-Minded as a Pole" (Concepts of Russian Common Consciousness "Germans" and "Poles" According to Russian National Corpus). *Studia Rossica Posnaniensia, 35*, 35-46.

Firbas, J. (1965). A note on transition proper in functional sentence analysis. *Philologica Pragensia, 8*, 170-176.

Frink, O. (1984). Cohesive Word Order in Russian. *Michigan Academician, 16*(3), 411-414.

Gracheva, V. (2013). *Markers of contrast in Russian: A corpus-based study.* (Master's thesis). University of Washington, Seattle, WA. Retrieved from ProQuest Dissertations and Theses. (AAI1542373)

Grenoble, L. A. (1998). *Deixis and Information Packaging in Russian Discourse.* Amsterdam: John Benjamins Publishing Company.

Halliday, M. (1967). Notes on Transitivity and Theme in English Part II. *Journal of Linguistics, 3*, 199-244.

Hentschel, G. (1992). On the Influence of the Order of Constituents on Choice of Case in Russian. *Lingua, 87*(3), 231-255.

Kallestinova, E. D. (2007). *Aspects of Word Order in Russian.* (Doctoral dissertation). University of Iowa, Iowa City, IA. Retrieved from ProQuest Dissertations and Theses. (AAI3271665)

King, T. H. (1993). *Configuring topic and focus in Russian.* (Doctoral dissertation). Stanford University, Stanford, CA. Retrieved from ProQuest Dissertations and Theses. (9403968)

Kizach, J. (2012). Evidence for weight effects in Russian. *Russian Linguistics, 36*(3), 251-270.

Kriesing, E. (1977). "Free" or "Restricted" Word Order in Russian. *Wissenschaftliche Zeitschrift der Pädagogischen Hochschule Karl Liebknecht, 21*(2), 231-235.

Krylova, O., & Khavronina, S. (1988). *Word Order in Russian.* Moscow: Russky Yazyk Publishers.

Lehmann, S. (1982). Word Order in Russian and the German Article. *Zielsprache Russisch, 2*, 43-49.

Levy, R., Federenko, E., & Gibson, E. (2013). The syntactic complexity of Russian relative clauses. *Journal of Memory and Language, 69*(4), 461-495.

Malamud, S. (2002). Centering in Russian. *University Pennsylvania Working Papers in Linguistics, 7*(2), 95-108.

McEnery, T. & Wilson, A. (2001). *Corpus Linguistics: An Introduction* (2nd ed.). Edinburgh: Edinburgh University Press

Mezhevich, I. (2001). Locative Inversion, Definiteness, and Free Word Order in Russian. *Calgary Working Papers in Linguistics, 23*(Spring), 30-48.

Nesset, T., & Makarova, A. (2012). 'Nu-drop' in Russian verbs: a corpus-based investigation of morphological variation and change. *Russian Linguistics, 36*(1), 41-63.

Pereltsvaig, A. (2004). Topic and Focus as linear notions: evidence from Italian and Russian. *Lingua, 114*, 325-344.

Sharoff, S. (2006). Methods and tools for the development of the Russian Reference Corpus. In A. Wilson, D. Archer, & P. Rayson, *Corpus Linguistics Around the World* (pp. 167-180). Amsterdam: Rodopi.

Smith, S. D. (2013). *Pro-Drop and Word-Order Variation in Brazilian Portuguese: A Corpus Study.* Brigham Young University, Provo, UT. Retrieved from BYU Electronic Theses and Dissertations. (Paper 3719)

Vintseler. (1977). Word Order and Context in the Russian Language. *Revue roumaine de linguistique, 22*(4), 489.

Yokoyama, O. (1986). *Discourse and Word Order.* Amsterdam: John Benjamins.

Zasorina, I. N. (1977). *Chastotny slovar' russkogo yazyka.* Moscow: Russkij Yazyk.