5-1-2011

The Role of Formulaic Language in the Creation of Grammar

Natalya Stukova

Natalya P. Stukova
Candidate

Linguistics
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Melissa Axelrod, Chairperson

Sherman Wilcox

Larry Gorbet

David W. Dinwoodie

THE ROLE OF FORMULAIC LANGUAGE IN THE CREATION OF GRAMMAR

BY

NATALYA P. STUKOVA

B.A., English, University of North Dakota, 1997
M.A., Linguistics, University of New Mexico, 2000

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Linguistics

The University of New Mexico
Albuquerque, New Mexico

May, 2011

Dedication

This work is dedicated to my parents, Pavel Stukov and Tamara Stukova.

I am forever grateful for your love and support.

ACKNOWLEDGMENTS

I am in debt to many people who have helped me with their ideas, encouragement, and support during the process of writing this dissertation.

I am most grateful to Melissa Axelrod, for being a great teacher and a great human being. Her encouragement and support at various stages of the project were indispensible. Without her input and inspiring help the completion of this project would not be possible. I am grateful to Joan Bybee for introducing me to the field of formulaic language, for her pioneering research in the field of usage based linguistics, and for her cutting-edge ideas and forefront inquiries into the nature of the human mind. I thank Sherman Wilcox, whose ideas go beyond the scope of linguistic topics per se and into a broader area of cognitive science, which always inform and challenge my own research and intellectual quest. I am grateful to Larry Gorbet for many discussions of linguistics over the years and for taking my every thought and interest, however small, seriously. I thank David Dinwoodie for his unique perspective on the linguistic issues presented here. His input enriched the project and provided the necessary angle to see the problems in a new light

I thank my many friends at UNM Department of Linguistics, in particular Melvetha Chee, Evan Ashworth, Michael Schwartz, Simoni Valadares, Valentina Kingsolver, and Maria Sotnikova for providing good humor and plenty of shop talk in the lab and beyond. I thank my many students in the introductory linguistic courses, whose questions were intriguing, and whose novel ways of looking at things were stimulating.

And finally, I thank my family, Tamara, Pavel, and Igor for their never ending love and support in good and in challenging times.

THE ROLE OF FORMULAIC LANGUAGE IN THE CREATION OF GRAMMAR

BY

NATALYA P. STUKOVA

ABSTRACT OF DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Linguistics

The University of New Mexico
Albuquerque, New Mexico

May, 2011

THE ROLE OF FORMULAIC LANGUAGE IN THE CREATION OF GRAMMAR

BY

NATALYA P. STUKOVA

B.A., English, University of North Dakota, 1997
M.A., Linguistics, University of New Mexico, 2000
Ph.D., Linguistics, University of New Mexico, 2011

ABSTRACT

Research in the field of Formulaic Language has shown it to be a very diverse phenomenon in both the form it takes and the functions it performs (e.g., Erma and Warren, 2000; Wray, 2002). The proposal made by Sinclair (1991) states that language as a system is organized according to two principles, "the idiom principle", which includes the use of all multi-word prefabricated sequences, and "the open choice principle," which covers word-for-word operations. Formulaic language is the embodiment of the idiom principle and constitutes the core of linguistic structure. Therefore, it must be subjected to scientific scrutiny from the variety of perspectives – typological, psycholinguistic, socio-pragmatic, and language acquisition.

This dissertation reports on the percentage of formulaic sequences - *prefabs* - in spoken and written Russian; the distribution of prefab types across two spoken and four written genres, and their interaction with non-prefabricated language and the impact that prefabs have on the structure of a particular language type. Russian is the language

typologically and structurally different from English. The main structural difference between English and Russian is that the Russian language has a free word order, wide inflectional system to code grammatical relations, and a satellite verb system. I hypothesize that these structural differences influence the quantity and the nature of formulaic sequences used in the language, the nature of alternation of prefabricated and non-prefabricated strings, and the preference of the speakers for one rather than the other aforementioned principles.

The method applied in the analysis of Russian prefabs is developed by Erman and Warren (2000) and originally was applied to the analysis of the English texts. This dissertation seeks to address a methodological issue of applying this method to typologically different languages. It has been argued (Garcia and Florimon van Putte, 1989) that the fixedness of the English word order contributes to the co-occurrence of elements and the formation of formulaic sequences in English. In this case, formulaic language becomes a language-specific tendency pertaining to English, and not a universal mechanism for language storage, processing, production and use.

The findings support the usage-based approaches driven by forces resulting from the frequency of use, discourse and communicative functions, grounded in the fine balance between the economy principle and the power of language creativity. The results of the study are used to draw implications for language processing and language modeling.

As we continue to perfect the methods of identification, classification and analysis of formulaic sequences, we will be in a better position to describe not only the amount but the nature of formulaic language, its interaction with non-formulas, and the impact

this alternation has on the linguistic structure as a whole. The current study investigates the nature of formulaic language in a free word order language. We seek to apply the method of identification, classification and analysis of prefabs, its interaction with each other and with non-formulaic language, as well as the estimation of choices made in producing spoken and written language.

My dissertation results suggest that a free word order language uses at least as many prefabs as a fixed word order language. On average, in a free word order language like Russian 65% of spoken and 58% of written language is composed of multiword formulaic sequences. The results strengthen the hypothesis that the idiom principle is a mechanism of global linguistic organization and processing. The proportion and distribution or prefabs is less affected by language type than by spoken written medium distinction and genre variation.

In addition, the results show that prefabs are frozen structures not amicable to standard syntactic transformations even in a free word order language. The results support the dual system of language processing, i.e., holistic and analytic, present in a free word order language.

TABLE OF CONTENTS

List of Figures

Chapter 2

Chapter 3

Chapter 5

List of Tables

Chapter 3

Chapter 4

Chapter 5

Chapter 1

INTRODUCTION

1.0 Introduction

This study investigates the nature of formulaicity in a free word order language and the role it plays in the grammatical structure of the linguistic system. Traditionally, formulaicity has been treated in linguistics as a marginal phenomenon reserved for such cases as idioms, collocations, clichés, proverbs, sayings, names, slogans, and other examples primarily from phraseology, which were supposedly outside of the core grammar of language. Formulaicity is a multifaceted phenomenon that can be characterized in multiple ways. Among other things, it includes such properties as semantic opaqueness, syntactic irregularity, and fixedness of word order, which historically have been viewed as problematic for the traditional account of language, which is based on the regularity of combinatorial rules, predictability of semantic composition, and the innateness of linguistic structures (Chomsky, 1957). It is widely accepted that idiomatic multiword sequences - due to their semantic or syntactic irregularity - are stored and retrieved whole from memory in a manner comparable to the way individual words of a language are stored and retrieved (Swinney and Cutler, 1979). Additionally, it has been suggested that multiword sequences that are completely regular and transparent also can be stored in the mental lexicon and accessed holistically (Langacker, 1987; Trambley et al., 2007). Regular multiword sequences can be formulaic due to the level of entrenchment in mental representations and their acquired conventional status. What makes these frequently co-occurring and non-idiomatic multiword combinations formulaic is the preference of native speakers for such

combinations over some others that are equally grammatical but nevertheless not preferred. If we take into consideration native-like selection, in addition to semantic and syntactic irregularity, the volume of language that can be termed formulaic becomes very large. None of these types of formulaicity, however, are amenable to traditional analysis. For this reason, formulaic language was relegated to the periphery, supposedly outside of grammatical structure to be studied and accounted for in linguistic theory.

The status of formulaic language has been challenged recently by a number of researchers (Sinclair, 1991; Pawley and Syder, 1983; Hopper, 1981; Barlow, 2000; Wray, 2002; Erman and Warren, 2000; Foster, 2001; Jackendoff, 1995, among others). Erman and Warren (2000) in particular demonstrate that formulaic language accounts for 58.6% of spoken English discourse that they analyzed and 52.3% of the written texts. Foster, by using a different method, shows that 32.3% of spoken data consists of formulaic language (Foster, 2001). Jackendoff (1995), based on the analysis of a small corpus study of television programs, argues that the number of formulaic sequences in certain genres may be even greater than the number of single words in the traditional lexicon. Mel'cuk (1995) ascribes even greater significance to formulaic language than the above mentioned authors. These recent demonstrations have drawn attention to the pervasiveness of formulaic sequences, highlighting the variety of forms they take and the functions they perform.

The statistics across studies may vary based on the sources of data and methods used in the identification, classification, and analysis of the formulaic sequences. However, this new body of research illustrates that the use of formulaic language is pervasive. Therefore, it cannot be relegated to linguistic periphery, but is, in fact, central

to the creation of global grammatical structure and to linguistic production and comprehension. The sheer amount of language that can be termed formulaic calls for a serious investigation of the phenomenon so that we can aim toward its modeling within a comprehensive theory of language.

Cross-linguistic studies of formulaicity are scarce to this day. All of the studies of formulaic language cited above have focused on the English language as the data source. English is a language that uses word order to signal grammatical relations in a sentence. Thus, English has a fixed word order that does not allow wide flexibility of ordering of elements within a clause. One might reasonably argue that fixedness of the elements in a clause contributes to the co-occurrence of elements and the formation of prefabricated strings. It would follow that the rigidity of the English syntax might contribute to the high percentage of prefabricated multiword sequences in the language as well as to the preference of English speakers for formulaicity in the production of utterances.

Based on this observation, I question whether the percentage of prefabricated strings in a language with a free word order is different from a language with a fixed word order. I hypothesize that the flexibility of the word order in a language provides its speakers with a greater number of choices in coding and decoding of utterances. We can expect a correlation between the amount of holistic processing and language type. Thus, the current study presents an analysis of the proportion, distribution, and length of formulaic sequences in a free word order language: Russian. The Russian language is an example of a conservative Indo-European system with rich inflectional and derivational morphology that provides support for the free word order syntax.

The examination of Russian and the comparison of Russian and English prefabricated language will reveal the interconnection between the flexibility of the word order, i.e. language type, and the mechanisms involved in language processing. The questions raised are: (a) Do speakers of various language types process linguistic structure differently (Also, do speakers store different units)? (b) Does the language type (i.e. linguistic structure of a language) influence processing? This study aims to demonstrate the role of formulaicity in a language typologically different from English. The English and Russian comparison provides the necessary contrast to study the nature and function of formulaic language itself and the role it plays in the universality of storage, processing, and structure of linguistic system.

An additional goal of this study is methodological: to investigate an application of a particular method cross-linguistically. This practical methodological issue, if resolved, may have a far-reaching effect by providing tools for more cross-linguistic comparison of multiple languages. If there is a way to compare languages with some degree of precision, this can enrich typological studies of language universals, enlarge our understanding of the workings of human language, and support the new view of grammar, which is emergent and usage-based.

1.1 What is Formulaic Language?

Formulaic language is a term used by many researchers to refer to the large units of processing – that is, lexical units that are more than one word long (Wray, 2002). Formulaic status of multiword sequences can be due to either their semantic or syntactic irregularity, or to the conventionalization of semantically and syntactically regular multiword sequences.

Historically, *formulaicity* has been the focus of study in phraseology (Winert, 1995) where it is equated with *idiomaticity*, and was avoided in generative accounts of grammar due to its presumably marginal status. Traditionally, idiomaticity is viewed as primarily a semantic phenomenon, though idioms have been characterized in many ways. Routledge Dictionary of Language and Linguistics (RDLL), 1996 defines *idiomatics* as "lexically or syntactically unchangeable groups of words that frequently have the value of a sentence and are thematized as a formula of politeness or greeting according to a pragmatic point of view." Idioms are frequently seen as semantically noncompositional where the meaning of the whole expression cannot be derived from the individual elements that constitute it. RDLL defines an idiom as "a multi-elemental group of words – or lexical entity with these characteristics:

(1) The complete meaning cannot be derived from the meaning of the individual elements.

(2) The substitution of single elements does not bring about a systematic change of meaning, which is not true of non-idiomatic syntagms.

(3) A literal reading results in a homophonic nonidiomatic variant, to which the first two conditions no longer apply".

Idiomatic, formulaic sequences such as these are conventionalized structures that exist due to their entrenchment as units. Langacker (1987) and other cognitive linguists argue that linguistic structures fall along a continuous scale of entrenchment in cognitive organization. Every use of a structure is thought to increase its degree of entrenchment.

> An event type has unit status when it is sufficiently well entrenched that it is easily evoked as an integrated whole, i.e., when it constitutes an established routine (Langacker 1987:100).

Thus, completely regular from semantic or syntactic point of view multiword sequences also can be entrenched because of the frequency of use and achieve a unit status. As a result, they become formulaic, which means processed holistically as a chunk.

1.2 Project Background

The basis for the current project is the study by Erman and Warren (2000) that seeks to discover the nature of formulaic language in spoken and written English discourse and the mechanisms behind its interaction with a non-formulaic language.

The empirical evidence based on the English data presented by Erman and Warren revealed that more that 55% of text consists of ready-made, multiword combinations of various length, which constitute a single retrieval for a language user, a prefabficated sequence or "prefab*"* for short. A prefabricated, multiword sequence is formulaic based on the fact that it is stored and retrieved as a whole. Thus, a formulaic or prefabricated sequence is a "chunk" of language that is processed, stored, and produced holistically as a single unit. The notion of holistic processing has a correlation with the notion of choice in psycholinguistics. If we take every single word of a text to represent a choice to be made by a speaker, the presence of prefabricated sequences reduces the number of choices to be made.

Erman and Warren (2000) developed a unique methodology to study prefabricated sequences and the interaction of holistic and analytic processing.
The method consists of several interrelated components. The first – prefab analysis -- seeks to document the number, type and proportion of prefabs in spoken and written texts. In addition to statistics, great effort is taken to differentiate between various types

of prefabs - lexical, grammatical, pragmatic, reducibles – to determine if various types of prefabs serve any particular main functions. The second – analysis of choices – tracks the number of lexical and overall choices made during language production. The presence of prefabs in any text means the reduction of choices made by the language user.

These authors analyze spoken and written sources separately due to the differences between the two modes. It has been determined that each medium is associated with distinct psycholinguistic and usage processes (Chafe, 1994). It is anticipated that prefab analysis will reveal variation between spoken and written medium in a free word order language as well.

The limitation of Erman and Warren (2000) study includes a lack of genre differentiation among written sources; all written texts are treated as a homogeneous entity and does not take into account possible variation in genres and styles. Likewise, the study does not report on variation between individual speakers or speaking genres. The current study proposes to differentiate between four written and two spoken genres. In addition, the variation among speaking tasks in the use of prefabs is also investigated.

1.3 Cognitive Grammar

1.3.1 Basic notions of cognitive grammar

The central claim of Cognitive Grammar (CG) as formulated by Langacker (1987, 1990, 1991, 2008) is that lexicon and grammar form a continuum consisting solely of assemblies of symbolic structures. A symbolic structure is the pairing of a semantic structure and a phonological structure. Grammar is this view is fundamentally symbolic, where conceptual structure is represented by means of sounds and gestures. Semantic structures consist of both conceptual content and the construal imposed on that content.

Langacker argues that the ability of the human mind to construe the same situation in alternate ways is reflected in linguistic structure. It comprises such factors as perspective adopted, the prominence accorded various elements, and characterization at a certain level of the scale of specificity/schematicity. Particularly relevant for our purposes are two kinds of prominence.

Profiling – within the extent of the conceptual content it evokes – its base – an expression directs attention to a particular substructure, called its profile, characterized as the entity the expression is construed as designating (its construal referent). Expressions evoking the same conceptual base can nonetheless differ in meaning by virtue of imposing different profiles on it. Langacker argues that an expression's grammatical class is not determined by its overall conceptual content, but specifically by the nature of its profile. A noun profiles a thing, abstractly defined. A verb profiles a process, defined as a relationship scanned sequentially in its evolution through time. Such classes as adjective, adverb, and prepositions profile relationships that are non-processual (atemporal in the sense that evolution through time is not in focus).

When a relationship is profiled, its participants are made prominent to varying degrees. The most prominent participant, called the trajectory (tr) is construed as the entity being located, evaluated or described. It is the primary focus ("figure") within the profiled relationship. Often another participant is made prominent as a secondary focus. This is called a landmark (lm).

Expressions can have the same content and profile the same relationship but differ in meaning due to contrasting choices of trajectory and landmark.

Having considered individual symbolic structures, I now turn to symbolic assemblies or constructions. First, a terminological point: my use of the term construction is broader than in Construction Grammar. As Langacker employs it, any symbolically complex expression – be it fixed or novel, regular or irregular – constitutes a construction. He also applies the term to any schematic pattern for assembling complex expressions (as well as a network of constructional variants).

In Cognitive Grammar there is no differentiation between regular and idiomatic multi-word sequences: both are symbolic assemblies which reflect a cline of idiomaticity. If some happen to be assembled according to a "rule" or captures by a schema and others are entirely idiosyncratic, all of these sequences are form/meaning pairings; Thus, they represent an inventory of constructions or constructional schemas of the a language.

1.4 The Study of Formulaic Sequences within Functional and Cognitive Grammar

The basic model adopted for the study is the one described by Alison Wray in Formulaic Language (2002.) The model is comprehensive in terms of accounting for regular as well as irregular phenomena in language. She proposes two mechanisms – analytic and holistic – for linguistic production and processing. In earlier introduced terminology, analytic processing corresponds to the open choice principle, which accounts for online generation of word-for-word utterances. Holistic processing, on the other hand, corresponds to a principle of idiom where the whole chunks of language of various length are processed as a unit.

In this model, formulaic sequences are viewed a dynamic response to processing and interactional needs of language users. Since formulaic sequences do not represent a

homogeneous set, but rather items of various strength and degree of fusion, performing various functions, some appear only for a short time while others become a permanent feature in a speaker's vocabulary. The model is in contrast with the traditional generative view of clear-cut boundaries between rule-generated strings and everything else that does not fit the pattern being discarded. The proof for psychological plausibility of the model comes from experimentation on native adult speakers, child language, second language speakers as well as aphasic patients (Wray, 2002).

1.5 Goals for the Study

The seminal study by Erman and Warren (2000) is based on the assumption that in language production the language user relies on the alternation between the use of word-for-word combinations – the open choice principle – and the use of preconstructed multi-word combinations – the idiom principle (after Sinclair, 1991.) The goal of their study was to show the impact that this alternation has on the language processing and production of texts. Erman and Warren refer to the written as well as to the spoken medium of language use as texts. The findings of their study demonstrated that more than half of any English text contains preconstructed multiword combinations. This finding convincingly demonstrated the pervasiveness of formulaic language and the reliance of the speakers on the holistic processing. The findings, while convincing, call for a cross-linguistic examination of the phenomenon, which will aid in our understanding of the mechanisms of language processing, storage and use.

This informative study is done on English, a language that has a fixed word order that does not allow much flexibility. One might argue that the rigidity of the English

syntax might account for the high percentage of prefabs in texts and this preference of English speakers for idiomatic principle in the construction of sentences.

In this case, idiomatic principle becomes a language-specific feature pertaining to English rather than a universal principle of language production. It is reasonable to think that the percentage of prefabs in language with the free word order might be different. We might think that the flexibility of the word order in a language provides its speakers with a much greater number of choices in coding and decoding of sentences. Thus, the speakers of these languages may have a greater tendency to use open choice principle in language production. This issue needs to be examined before conclusive remarks are made on the universality and interaction of such structural principles as open-choice and idiom principles.

In this light, the primary goal of the current study is to conduct a comparable test on a typologically different language, i.e., a language with a free or maximally flexible word order. Thus, the goal is to compare the average proportion of prefabs and average number of choices in a text of typologically different languages. This examination will reveal the interconnection between the flexibility of word order and the mechanisms involved in language production. Thus, it will demonstrate the role of formulaicity in a language typologically different from English and will get us closer to the understanding of the nature and function of idiom principle.

An additional methodological goal of this study is to explore the possibility of applying Erman and Warren's method cross-linguistically, specifically to languages that are typologically different. This methodological development is a key issue in our ability to compare very different linguistic systems on the basis of a common ground, i.e., a

common method of analysis. If there is a way to compare languages with some degree of precision, this can enrich typological studies of language universals, enlarge our understanding of the workings of human mind, and support the view of grammar that is grounded in empirical evidence its psychological plausibility.

1.5.1 Specific research questions

This study aims to investigate the types, frequency, and distribution of formulaic language in Russian, a free word order language. The current study uses the method of analysis developed by Erman and Warren (2000) for the English language, a fixed word order system. The following research questions will be investigated in the Russian data:

a. How many prefabs occur in spoken and written texts?
b. What are the prefab types and their distribution in a free word order language?
c. What is the average length of prefabs and non-prefabs?
d. How many choices are made in producing spoken and written Russian?
e. Is it significant if the discourse is written or spoken for the use of prefabs?

The English data provides some information in regards to these questions. Erman and Warren (2000) report for both spoken and written English discourse the number, type, and proportion of prefabs in texts, the distribution of various types of prefabs, the number of choices, both lexical and total, made by a language user, and the average length of formulaic and non-formulaic sequences in various texts.

The results of the English-based study demonstrate that 1) more than half of the analyzed data (spoken and written texts) consisted of prefabricated sequences; 2) prefabs are very heterogeneous in form and function and are distributed unequally; 3) the presence of prefabricated strings significantly reduces the number of choices made in the

production of a text, thus significantly reducing the cognitive load for language production and use; 4) There are no English texts, written or spoken, that consist exclusively of prefabricated strings or texts that have no prefabricated strings at all; rather the mixture of the two create the tapestry of linguistic structure. We can anticipate a similar interaction of the two mechanisms - formulaic and non-formulaic - in the Russian data, as the two mechanisms are rooted and are reflective of a deeper cognitive organization and a commonality of social interactions across cultures. The analyses of the Russian data are presented in Chapter 4.

1.5.1.1 Quantity and frequency

Investigating the quantity and frequency of prefabs in Russian will allow us not only to document the nature and number of prefabs in a free word order language but to detect any influences of the free word order syntax on the processes of formulaicity. The descriptive work will lead us to ask important questions relevant to the cognitive mechanisms involved in language storage, production, and use. Among them will be, for example, such questions as a) Is the alternation between prefabricated and non-prefabricated language a language-specific or a universal phenomenon? b) Do speakers of typologically different languages rely equally on the idiom and open choice principles? c) Does it make any difference if discourse is spoken and written? One additional question not been considered by Erman and Warren is whether the number of prefabs varies among genres of written texts or among speakers of the same dialect. The study of Russian prefabs investigates four types of written and two types of spoken texts to determine the possible variation across genres. Based on the empirical evidence, we will

be able to make comparisons with the English data and determine if flexibility of the word order affects the formation of prefabs and prefab types.

1.5.1.2 Types and distribution of prefabs

According to Erman and Warren (2000: 38) typology, prefabricated sequences may be classified into four types, which will be more thoroughly described in Chapter 3:

a. Lexical prefabs, or referring items, e.g., *subject matter, a waste of time, be in touch with.*
b. Grammatical prefabs, or text-forming items, e.g., *for example, be going to, and so on.*
c. Pragmatic prefabs, which provide textual, interpersonal, and metalinguistic stance, e.g., *I mean, you know, and everything.*
d. Reducibles: I'll, she's , we're, etc.

Various types of prefabs serve different purposes in language. The existence of lexical, pragmatic or grammatical prefabs reflects different reasons for conventionalization of phrases. Lexical prefabs are semantic units that represent extra-linguistic entities or phenomena. The notional characterization of lexical prefabs is based on the fact that lexical prefabs denote entities, properties, states, events, situations, places, positions, and periods of time – *out of date, run off, here and there, permanent job, subject matter, at the time, by then*. Lexical prefabs are believed to exist because there are standard ways of referring to standard situations and phenomena in a culture. Idiomaticity arises precisely because standard situations naturally tend to be expressed in familiar ways.

Functional prefabs comprise grammatical and pragmatic prefabs. Grammatical prefabs are intralinguistic text-forming items rather than units with extralinguistic reference. Typical examples of grammatical prefabs are determiners, quantifiers, proforms, tense, aspect, and mood-forming elements, links and intensifiers. Grammatical prefabs are considered text forming rather than referring units. That is to say, they will quantify, specify or modify the reference or meaning of nouns, verbs, adjectives or adverbs in a general matter, or they will serve as their substitutes or as links between propositional or referring items. Again, the grammatical prefabs exist because there are standard ways of monitoring texts in language.

Pragmatic prefabs text-monitoring units which do not directly partake in the propositional content of the utterances. They may occur outside of the syntactic structure, and that way they are different from grammatical prefabs. Brinton (1996: 38) suggests that there are two main categories of pragmatic markers: textual – *and then, I mean, well you know*, and interpersonal – *I see, I think so, what's the word.* Erman and Warren add a third component to this classification – metalinguistic monitors – *and everything, sort of , I should think, I must say*. Pragmatic prefabs exist because there are standard ways of expressing oneself in standard social interactions that are recognized by the community members as natural and native-like.

Reducibles are the last category of Erman and Warren's classification. Reducibles are types of abbreviation common in English that fall into the four main groups: 1. Pronouns and auxiliary - *I'm, he's, we're;* 2. Auxiliaries – *don't, isn't, hasn't, can't;* 3. Auxiliary and auxiliary - *would've, should've;* 4. Let + us: *let's.* It can be easily argued that these combinations of items represent units of language that are processed and

produced holistically as a single chunk. The English orthography reflects the actual usage events in articulation and processing.

While articulatory and cognitive processes that give rise to the phenomenon of chunking are part of natural language use, they are usually not prescriptively regulated or controlled. Orthography, on the other hand, is often subject to such regulations. Reducibles can be viewed as conventions of the English spelling and a language-specific phenomenon. While these natural processes of articulatory fusion and chunking take place in all languages, not all orthographic conventions would necessarily reflect that. The Russian language, for example, does not use this type of abbreviation so common in English. For this reason, the category "reducibles" is not used in the present study in the analysis of Russian prefabs. The current analysis of Russian formulaic language adopts a three-way classification of prefabs: lexical, grammatical, and pragmatic.

1.5.1.3 Length of prefabs

One of the research questions (1c) considers the length of prefabs and non-prefabs in spoken and written discourse. The specific question is the average and the maximum length of prefabs and non-prefabs in various text types.

In general, the length of prefabs tends to be shorter in spoken language than in written. This is expected in light of the differences between the production of speech and writing. Thus, the average length of spoken prefabs in English is 2.61 words/prefab while in Russian it is 2.58 words/prefab. The average length of written prefabs is 2.80 words/prefab in English and 2.87 in Russian. The longest prefab in English in Erman and Warren data is 14 words/prefab, and in the Russian pilot study it is 12 words/prefab. The data is currently available for various types of English prefabs. Thus, lexical prefabs are

longer than other types of prefabs - 3.03 words/prefab, whereas grammatical prefabs on average are 2.26 words/prefab; pragmatic, 2.26 words/prefab; and reducibles are 2.00. The Russian data is presented in Chapter 3 for the pilot study and Chapter 4 for the current study, and both are discussed fully in Chapter 5.

1.5.1.4 Choice

The second component of the method includes the "analysis of choices", which is a tracking system of the number of lexical choices as well as the overall number of choices made by speakers or writers in language production. Since prefabricated units comprise at least two words and are used as a whole chunk of language, they reduce the number of choices the speaker or writer needs to make in producing the text. In addition, all choices are not considered equal. A lexical choice often represents a choice from an unlimited list whereas a choice of a particular grammatical form is often limited by the categories present in a particular language. Thus, unrestricted lexical choices represent a heavier cognitive load compared to choices of grammatical or even pragmatic forms. The number of choices made in language production is indicative of the cognitive effort expended by a language user during the language production. The analysis of prefabs and choices combined form the basis of the Erman and Warren's method of studying formulaic language qualitatively and provide very interesting results for the English data.

The question of choice is particularly interesting when comparing Russian and English systems because it reveals the fine-grained mechanisms of linguistic coding and the cognitive cost that it takes to produce various types of languages.

1.5.1.5 Summary

As described above, the current analysis proceeds in the following steps. First, the number, types and distribution of prefabs is established in various genres of spoken and written discourse. Second, the number of single choices made in the production of texts is calculated. The distinction between lexical and non-lexical choices is significant and, therefore, counted separately. Lexical choices represent a heavier cognitive load because any lexical choice is made from unlimited options, whereas grammatical or pragmatic choices, for example, are made from a limited set of options offered by any particular language. Third, the length of prefabs – average and maximum – is considered in order to have a visual representation of the alternation of formulaic and non-formulaic material. Additional considerations include the functionality of prefabs in a linguistic system: a) Do prefabs serve certain main functions? b) How and why can prefabs be varied? c) Which prefabs are stylistically neutral, if any? These are some of the questions the current study proposes to answer, specifically in a free word order language. The results, however, go beyond language-specific findings. The comparison of two distinct linguistic systems reveals new insights for the understanding of language.

1.6 Organization of the Study

Chapter 1 is an overview of the dissertation. It introduces the key concepts of idiomaticity, formulaicity, prefabricated expressions. It covers the project background, objectives of the current study, and the methodology used.

Chapter 2 describes treatment of formulaicity in various theories of language – formal, functional, and cognitive. The field of phraseology, which originally investigated the phenomenon of formulaic language, equates formulaicity with idiomaticity.

Traditionally, idiomaticity is viewed as an instance of "dead" metaphors, whereas functional and cognitive theories view a large portion of idiomatic expressions as transparent, structured and metaphorically motivated. They draw on the knowledge of the conceptual system, which is largely metaphoric in nature, in explaining the meaning and structure of idioms. Chapter 2 also addresses a number of key topics such as the organization of language in terms of lexicon vs. syntax (Section 2.3.1), an issue of access, processing, and choice (Section 2.5), and word order typology (Section 3).

Chapter 3 presents a detailed account of the method of prefab analysis developed by Erman and Warren (2000) and the analysis of choices (Section 2). Modifications of the method for the Russian data are described concurrently. In this chapter, types of prefabs are discussed in detail in Sections 2.2.1. through 2.2.4. The pilot study conducted earlier in the course of this research is presented in Section 3. It describes written and spoken corpora used in the study, gives comparison of English and Russian data, and presents preliminary results and conclusions. The data and coding for the current study is presented in Section 4. Both spoken and written corpora have been expanded significantly and are used as a data for the current study. The current study also codes for various text types that represent different spoken and written genres. Spoken corpora include recordings of live interviews as well as naturally occurring informal conversations. Written discourse is represented by four distinct genres described in Section 5.2.

Chapter 4 presents the results of the current study. It reports on the number, type, distribution of prefabs in Russian texts in Section 1. The number of lexical and overall

choices are presented in Section. The length of various types of prefabs – average and maximum – is presented in Section 2.

Chapter 5 presents discussion of the results of the current study. It compares the Russian data with the English data from Erman and Warren (2000), on which the current study is based.

Chapter 6 describes overall conclusions of the study and gives direction for future research.

1.7 Summary of Chapter 1

Chapter 1 is an overview of the current study. It introduces the key concepts of formulaic language, idiomaticity, prefabs, and others. This chapter gives the background of the current project, introduces the method of analysis, and presents the design of the current study. The study of formulaic sequences within functional and cognitive grammar is discussed. The framework of analysis adopted for the current research is cognitive grammar. The goals of the study and the specific research questions are described in Section 5. These include the analysis of prefabs and the analysis of choices in the Russian language. Russian, being a free word order language, is used as the contrast to the English language, which is considered to be a fixed word order language. The analysis of prefabs covers issues of frequency (Section 5.1.1.), types and distribution of prefabs (Section 5.1.2), average and maximal length of prefabricated expressions (Section 5.1.3), and the analysis of choices (Section 5.1.4). The general organization of the study is described in Section 6.

Chapter 2

Theoretical Background

2.0 What is Formulaicity?

Formulaic language is defined as "multiword collocations which are stored and retrieved holistically rather than being generated *de novo* with each use" (Milwaukee Symposium on Formulaic Language, 2007.) Linguists have long recognized the existence of formulaic language and, to some extent, of recurrent patterns of language (see Wray, 2002, for an overview). With the advent of the Chomskian paradigm, the scientific study of formulaic language was put on hold. The generative tradition has long proclaimed the autonomy of grammar from meaning and has put semantics and usage considerations out of the linguistic theory proper and did not afford formulaicity its rightful place. Not being able to incorporate various types of irregularity associated with the formulaic language, generative models chose to minimize the scope of the phenomenon or, at times, discard it all together.

Formulaicity, being such a common feature of natural language, could not have been left unattended for too long. In recent years, research in formulaic language has grown to include an impressive number of domains, including language acquisition (Wong-Fillmore, 1976; Peters, 1983; Granger, 1996; Weinert, 1995), language modeling (Wray, 2002; Langacker, 1988), natural languages processing (MacWhinney, 1997), conversational routines (Aijmer, 1996; Coulmas, 1981), probobalistic strings and collocations (Juravsky, Alternberg, 1996 a, b; Kjellmer, 1994; Sinclair and Renouf, 1991) and language pathology (Wray, 2002).

As a field of study, formulaic language is handicapped by a bewildering array of variously defined terms (Wray and Perkins, 2002). Partly because of the uncontrolled and non-standardized terminology, classification of formulaic expressions is difficult. Wray (2002) reports more than 40 different terms used to refer to formulaicity. A snapshot of terms used to describe formulaic language in different frameworks is given in Figure 1 (adopted from Wray, 2002).

amalgams – automatic – chunks – cliches – co-ordinate -
constructions – collocations – complex lexemes – composites -
conventionalized forms – fixed expressions – idioms -
formulaic language – formulaic speech –
formulas/formulae – fossilized forms – frozen metaphors – frozen
phrases – gambits – gestalt – holistic – holophrases – idiomatic –
irregular – lexical simplex – lexical(ized) phrases –
lexicalized sentence stems – listemes – multiword items/units –
multiword lexical phenomena - noncompositional –
noncomputaional – phrasemes – praxons – preassembled speech –
recoded conventionalized routines – prefabricated routines and patterns –
ready-made expressions – ready-made utterances - recurring utterances –
rote – routine formulae – schemata – semipreconstructed phrases that
constitue single choices – sentence builders – set phrases – stable and
familiar expressions with specialized subsenses – stereotyped phrases –
stereotypes – stock utterances – synthetic – unanalyzed chunks
of speech - unanalyzed multiword chunks - units

Figure 1. Terms used to describe aspects of formulaicity (Wray 2002)

This terminological mine has caused confusion within formulaic language research. It is not uncommon when researchers within the same discipline use different labels for the same phenomenon, as well as the same label for different phenomena. The lack of uniformed terminology makes a comparison of studies and research findings a challenge.

Choosing a term for the unit of analysis for the current study was not easy. Several possibilities were considered. First, *formulaic sequence* (FS) is a frequently used term that covers a wide range of data from very obvious instances of formulaic language such as idioms, proverbs and sayings to more obscure and easily missed collocations that appear to be formulaic often only on a second examination. This term could have worked for our purposes except its all-encompassing nature makes it less suitable for the specific range of phenomena investigated in this study. Second, *formula* is frequently used, but it is less suited as a cover term since it is applied to instances with idiosyncratic conditions of use. Similarly, *lexical phrase* is used by Nattinger and DeCarrico (1992) to emphasize the relationship between formulaic language and functional language use.

In the present study, the terms *prefab* was chosen as a basic unit of analysis. Prefab is an abbreviation for prefabricated expression, a term gaining momentum due to the seminal study by Erman and Warren (2000) and subsequent studies (Bybee and Torres 2009). I follow the suggestion of Erman and Warren (2000) to define a prefab as:

> …. a combination of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization. (Erman & Warren 2000:31).

Thus, prefabs are high frequency, multiword strings that may span phrasal boundaries (e.g., *to the best of my knowledge*, *I don't know whether, don't worry about it*).

This definition is based on the preference for a particular combination due to conventionalization, which plays a paramount role in nativelike selection and nativelike fluency. Given their high-frequency of use, many researchers have argued that such multiword strings are stored and processed as single units.

Nattiger and DeCarrico (1993) point to some possible explanatory factors of this preference:

> …Lexical phrases are chunks of language of varying length, conventionalized structures that occur more frequently and have more idiomatically determined meaning than language that is put together each time.

Two stipulations can be induced from this definition. The first is that prefabs are conventional structures that occur more frequently, and the second is that there are at least two types of language: the one that is "put together each time" it is used and the one that is not. These two aspects deserve special attention and are discussed in more details later.

Another informative definition of a formulaic sequence is provided by Wray (2002), which was cited earlier in Chapter 1, Section 2 and is repeated here for convenience:

> A sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.

This definition echoes Nattiger and DeCarrico's one in terms of posting two types of mechanisms for processing. At the same time, Wray's definition points out that formulaic language is not only frequent and more idiomatic but stored and retrieved whole as opposed to language, which is generated and analyzed. Thus, words and word strings that appear to be processed without recourse to their composition are called formulaic (Wray, 2002:4). I return to the issues of storage and processing in later sections.

Despite the convergent definitions, identification of prefabs in practice is a challenging task. Erman and Warren (ibid: 33) point out several reasons why

identification of prefabs is difficult. First, due to variability that exists among speakers in a community, a chunk may or may not gain a status of a prefab, depending on a speaker. Some prefabs are known to all members of a linguistic community, whereas others are less prominent. The entrenchment of prefabs varies among speakers at a synchronic point of time. Prefabs may also vary in strength of entrenchment over a period of time in the same speaker. Conventionalization is a gradual process, where co-occurrence and fusion of elements in a prefab is a matter of a degree and time. Second, the identification of prefabs is difficult due to the transparent nature of some collocations. Such prefabs are not easily seen as non-compositional or idiomatic. Erman and Warren identify idioms, compounds that are spelled as separate words, non-compositional habitual collocations, and prepositional and phrasal verbs as indisputable examples of prefabs. Nevertheless, based on the findings of their study they conclude that "the identification of all and only the prefabs in a text is in practice impossible" (ibid: 33.) These researchers posit three criteria for identification of prefabs, which are adopted for the present study as well and are listed in Figure 2:

1. A prefab must be a combination of words (orthographically represented as separate words; e.g. *tea cup,* not teacup.

2. The combinations must manifest some feature of conventionalization. (used as evidence, but not proof that a combination is memorized)

3. Restricted exchangeability. At least one member of a prefab cannot be replaced by a synonymous item without causing a change of meaning or function and/or idiomaticity.

Figure 2. Criteria for Identification of Prefabs (Erman and Warren, 2000)

Erman and Warren identify four different types of prefabs: lexical, grammatical, pragmatic, and reducibles. Different types of prefabs serve different discourse-pragmatic and communicative functions. The specific reasons for conventionalization of different phrases will be explored in the course of this study.

The difference in the number of prefabs in written and spoken texts is close to 6%, which is much smaller than expected. The greater difference between written and spoken texts is found in utilization of particular types of prefabs. Lexical prefabs are used almost twice as much in written texts as in speech (71.5% vs. 38.8%); grammatical prefabs are used slightly more in speech (20.5% vs. 16.9%); whereas pragmatic and reducibles are used significantly more often in speech (16.7% vs. 2.4% and 24% vs. 9.2% respectively).

This study, among other things, demonstrates that formulaicity is far from being a marginal phenomenon. It is not a matter of linguistic style as previously thought but a strong tendency that serves particular cognitive, discourse and communicative functions. Moreover, new empirical evidence shows that formulaic language is a vital mechanism that plays a role in the construction of grammar itself. Even though the nature and functions of formulaic language remain an open research question, one thing is clear: the pervasiveness of formulaic language is paramount. Therefore, it needs to be explained and integrated into the comprehensive theory of language.

Erman and Warren (2000) classify all prefabs into four categories: *lexical, grammatical, pragmatic* and *reducible*. Lexical prefabs refer to the extra-linguistic entities such as things, events, states, and situations. For the reason that lexical prefabs often have a real-world referent, they are easier to identify and to classify. The functional

prefabs, on the other hand, which include grammatical and pragmatic prefabs are intra-linguistic entities and serve various functions depending on their nature and use.

The existence of various types of prefabs reflects different reasons for conventionalization of phrases. This classification was modified for this study to accommodate the analysis of prefabs in the Russian language. The present study seeks to use Erman and Warren's (2000) classification as much as possible, so the results can be compared. Some modifications were necessitated by the differences that lie in the structure of the Russian language. My account of the formulaic sequences in Russian involves an analysis of three independent categories: *lexical, grammatical, and pragmatic* prefabs.

2.1 Idiomaticity

A range of different things in a language might reasonably be said to be idiomatic. Let us survey the kinds of phenomena that come under idioms in a broader sense.

First, the idiomatic character of an expression resides it its semantic value. The syntax of these expressions is unexceptional. What makes them idiomatic is that their meanings cannot be predicted from the meanings that the component words have elsewhere in the language. As a consequence, the expressions are open to two kinds of interpretation – the literal interpretation and the idiomatic interpretation. The latter has to be learned as a specific property of each of the expressions. We can distinguish two respects in which an expression can be regarded as idiomatic. First, we can have idiomaticity of meaning.

An expression is idiomatic to the extent that its meaning cannot be determined by the meanings of its parts and the manner in which they are combined. Here are examples

of lexical idioms: red herrings. Syntactically these expressions are perfectly regular and unexceptional. If we take the first one "red herring" the expressions is not created by assembling its components in accordance with a syntactic rule. This is because its meaning cannot be derived from the meanings of its components. The expression 'red herring' has to be listed as an item in a person's knowledge of their language. Since the lexicon is taken to be the domain of idiosyncratic; the expression will have to be included in the lexicon. Doing so, however, upsets the neat compartmentalization of lexicon and syntax, since from the syntactic point of view, red herring is unexceptional.

Second, we can have idiomaticity of form. Unlike non-idiomatic expressions, idioms are not freely assembled by the application of general syntactic rules. For a large number of expressions, the idiomaticity resides not so much in the special meaning but in the formal aspects of the expressions. The syntax may be somewhat irregular in comparison with other patterns prevailing in the language.

Or the idiomaticity may reside in collocational requirement that is not fully predictable from general principles. Idioms of this kind do not normally allow a literal and an idiomatic interpretation. In fact, the meanings of this kind of idiom are often fairly transparent and could be guessed at even by a learner who is encountering the expression for the first time. Nevertheless, because of their formal properties the expressions must be learned as such. A multiword expression may be conventionally associated with certain kinds of situations. Other expressions have a distinctive discourse-structuring function. Examples include:

A further group comprises conventionalized ways of expressing speaker attitude. Other examples of chunks of language that speakers commit to memory include catchphrases, clichés, proverbs, sayings, aphorisms, and parts of literary texts.

The last group of idioms that I want to mention is what Makkai (1972) calls "idioms of encoding" in contrast to "idioms of decoding." Idioms of encoding view the matter form the perspective of the speaker. Idioms of encoding have to do with the speaker's knowledge of the conventionalized way of saying something. A speaker not only needs to know the conventional meaning of an idiom, but also the circumstances in which the expression can be appropriately used. A focus on encoding greatly expands the range of things that appropriately can be said to be idiomatic. This is because the conventionalized way of saying something may not in itself be in any way idiomatic, from a decoding perspective. The idiomaticity resides in the fact that this happens to be the conventional way to say something, rather than some other equally plausible way.

Once the category of idioms of encoding is recognized, idiomaticity takes over quite a large part of the lexicon in that the conventionalized way of naming an entity, or a type of entity, needs to be regarded as idiomatic. It is clear that a very wide range of phenomena come under the rubric of the idiomatic, in a broad sense of the term. Fillmore, et al. (1988) argue that "the realm of idiomaticity in a language includes a great deal that is productive, highly structured, and worthy of serious grammatical investigation." The model of grammar that Fillmore and colleagues propose includes principles whereby a language can associate semantic and pragmatic interpretation principles with syntactic configurations larger and more complex that those definable by means of single phrase structure rules.

2.2 Generative Grammar

The theoretical positions least sympathetic to formulaicity as a principle feature of language structure, are the ones which propose a single grammatically based processing system. Human language has been characterized by generative grammarians as a homogeneous class, composed of definable discrete elements, which are combined according to specifiable grammatical rules. Chomskian paradigm that dominated the second half of the 20th century traditionally avoided idiomaticity due to its irregularity and non-compositionality. Universal Grammar (UG) viewed idiomaticity as a marginal and not significant part of human language due primarily to its inability to account for the phenomenon and not to any empirical or experimental evidence. In UG, the knowledge of language is comprised of the knowledge of the lexicon, a mental dictionary with lists of lexemes, and the knowledge of the combinatorial rules, which are part of the autonomous syntax.

The division between lexicon and syntactic rules in this model is clear cut, handled by different modules that do not overlap to any extent. The separateness of the lexicon and generative rules is justified by the theoretical necessity to explain linguistic creativity exhibited by native speakers. The traditional lexicon that consists of individual lexical items, listed and stored as separate representations, is viewed as a repository of everything idiosyncratic and irregular. The traditional lexicon combined with a large set of syntactic rules, according to the theory, can generate an infinite number of grammatical sentences in a language. Thus, the separateness of lists and rules, later identified as *a list/rule fallacy* (Langacker, 1987) requires an assembly of lexical items via generative rules every time the language is coded and decoded. The issue of

processing cost of such mechanism has never been addressed, and the lack of psycholinguistic evidence undermined the probability of such a mechanism. According to the mantra of the day, this generative mechanism provides its speakers with an ability to understand and produce sentences that have never been uttered. Thus, linguistic creativity, which is a hallmark of human language, was explained by the generative mechanism. However attractive this analysis was at the time, it was not based on empirical evidence and did not have any claim to psychological reality. Moreover, we can see how any linguistic phenomenon that is not describable in terms of lists and rules has to be minimized or discarded in generative grammar. This is largely why formulaicity has been overlooked for a good part of the 20th century in American structuralist linguistics.

The phrase structure (PS) representation in syntax which is central to generative theory formalism was presented as a universal organization of the linguistic system. However, even though this may sound very Whorfian, I am fairly sure that PS-syntax could not have been invented and developed by a native speaker of Latin or Russian. Such languages have incredibly flexible word order and very rich systems of morphological markings: word arrangements and inflectional affixes are obviously contingent upon relations between word-forms rather than upon constituency. English, on the other hand, promotes PS-representation in syntax with its rigid word order and almost total lack of syntactically driven morphology.

English-- in contrast to Russian -- is very exotic in that it uses constituency almost as its only expressive device in syntax, i.e., as the only device for encoding syntactic structure in actual sentences. Thus, constituency, which is marked by word order and

prosody, is the principle observable phenomenon used to indicate the underlying syntactic relations. As we know, constituency is a manifestation of syntactic structure, not syntactic structure itself. The pitfall of generative grammar is that it mistakes this idiosyncratic trait of English for a universal mechanism of syntactic representation. It is, then, to be expected that in languages such as Russian that do not use constituency as the main expressive device, the constituency, or phrase-structure formalism will perform poorly as a means to represent syntactic organization of language.

We can notice that dependencies between words that reflect syntactic relations are not a surface-observable phenomenon at all and therefore they are not language-specific. In Russian such syntactic relations of cause are symbolized by rich morphological system rather than by fixed word order.

Dissatisfaction with the traditional generative account leads a new generation of linguists to ask new questions: Do speakers really have separate boxes for lexicon and a set of combinatorial rules? What is an entry in the lexicon: a word, a phrase, a sentence? What is the capacity for memory storage of linguistic units? What is the processing cost of this generative mechanism? And ultimately, is this a psychologically plausible model of language?

2.3 Functional Grammar

2.3.1 Basic description of the model

The basic model adopted for the study is the one described by Alison Wray in Formulaic Language (2002.) The model is comprehensive in terms of accounting for regular as well as irregular phenomena in language. Wray proposes two mechanisms – *analytic* and *holistic* – for linguistic production and processing. In earlier introduced

terminology, analytic processing corresponds to the open choice principle, which accounts for on-line generation of word-for-word utterances. Holistic processing corresponds to a principle of idiom where whole chunks of language of various length are processed as a unit.

In this model, formulaic sequences are viewed as a dynamic response to processing and interactional needs of languages users. Formulaic sequences do not represent a homogeneous set but rather items of various strength and various degree of fusion that fulfill a variety of functions. Some appear only for a short time while others become a permanent feature in a speaker's vocabulary. The model is in contrast with the traditional generative view of clear-cut boundaries between rule-generated strings and multiword units processed as single chunks. The proof for psychological plausibility of the model comes from experimentation and empirical evidence gathered from various domains – language of native adult speakers, child language, second-language speakers, and aphasic patients (Wray 2002).

In Wray's (2002) model formulaic language is viewed as central to the grammatical knowledge of language users and can fulfill a plethora of functions. In all cases, the speaker could directly benefit from using prefabricated material. Many functions appear to aid the speaker, but some are partly, or even exclusively, are in the interests of the hearer.

Three functions of formulaic language that Wray (2002) identifies as primary in her model relate to issues of processing, identity, and manipulation. First, in the realm of processing, formulaic language is viewed as a means of reducing the processing load and

alleviating the speaker's processing effort. In many cases, formulaic sequences enhance the fluency of the speaker's output.

Formulaicity can be present when there are no pressures on the speaker's production if there are particular pressures on the hearer's comprehension, such as weather forecast, auctions, and military commands. Much stylistic repetition is dedicated to aiding the hearer's decoding by directing attention and reinforcing particular aspects in the content. Formulaic discourse markers seem able to support both the speaker's and the hearer's processing simultaneously. By mapping out the structure of the text, they help the speaker to remain focused, while making the content and the speaker's intentions easier for the hearer to follow.

Second, formulaic language signals speaker's identity as an individual and as a member of a group: (a) speakers seem to be able to express their identity as an individual using deliberately memorized strings and stylistic markers, (b) speakers can express their group identity by adopting customary ritualistic utterances, idiomatic turns of phrase, and collocations. Thus, formulaic material plays a central role in maintaining the identity of the community. The third function identified by Wray as primary is "manipulation" – formulaic language plays a role in manipulating the hearer into a desired action or perception. The motivation behind the desire to speak fluently, express identity, organize text and help the hearer to understand what you say is based on the speaker's desire to communicate effectively. This motivation ceases to view formulaic sequences as the solution to linguistic problems at all. Instead, it views them, in all of their uses, as linguistic solutions to a single, non-linguistic problem -- promotion of self (ibid: 94).

The questions arise: Why should the language that performs these functions characteristically be formulaic? (Why does the army use commands? Why do we say "excuse me" and leave quietly?) What would be the effect of not using formulaic sequences in these situations? Bernstein (1972) observes that "meanings which are discreet to the speaker must be offered so that they are intelligible to the listener."

According to Wray (2000), the speaker, by virtue of deciding what is expressed and how, can exert control over the range of linguistic interpretations open to the hearer and can make sure the intended message is received and interpreted in a desired manner. Wray (ibid: 35) claims "the more novel our output is for the hearer, the more likely it is to be misunderstood". Thus, the use of prefabricated material potentially guarantees less misinterpretation and misunderstanding. MacKay (1951) states that "successful communication depends on symbols having significance for the receiver, and hence on them being already in some sense prefabricated for him." With the use of novel utterances (e.g., not commands in the army), the potential for misunderstanding and for slow reactions is greatly increased. In a life-or-death situations (military orders), simple commands, previously learned by everyone and easily differentiated, are the most effective way of gaining a fast and uniform response from a large group. Wray argues that the goal of commands is the accurate manipulation of many people at the same time, and this is achieved by the use of word strings that are formulaic for the hearer. On closer consideration, it becomes clear that all of the functions of formulaic sequences serve a single goal: the promotion of the speaker's interests. These interests Wray, (ibid: 36) argues, include:

a) having easy access to information; b) expressing information fluently; c) being listened to and taken seriously; d) having physical and emotional needs satisfactorily and promptly met; e) being provided with information when required; f) being perceived as important as an individual; g) being perceived as a full member of whichever groups and deemed desirable.

It is in the speaker's interests to ensure that the hearer understands a message, because the intended effect of the utterance is to create a situation beneficial to the speaker. Bruner (1983) puts it this way: "Whether we request information, goods, services, or merely recognition we must accommodate the hearer's capacities, his constraints, our relation to him, and the conventions to which he adheres both in language and in the real world." The object of a request is to get somebody to deliver the goods. (Bruner 1983: 91) In the true spirit of functional grammar, the central tenet of Wray's model is that the structure of the language directly reflects and embodies the functions that language serves.

2.4 Cognitive Grammar

The conception of grammar adapted for this study is usage-based, specifically Cognitive Grammar (CG) as formulated by Langacker (1987, 1991). The three major tenets of CG are given here and serve as the primary conceptual framework for the present study.

1. Semantics structure is not universal; it is language-specific to a considerable degree. Further, semantic structure is based on conventional imagery and is characterized relative to knowledge structures.

2. Grammar (or syntax) does not constitute an autonomous formal level of representation. Instead, grammar is symbolic in nature, consisting in the conventional symbolization of semantic structure.
3. There is no meaningful distinction between grammar and lexicon. Lexicon morphology and syntax form a continuum of symbolic structures, which differ along various parameters but can be divided into separate components only arbitrarily.

Within the framework of CG, the symbolic nature of language extends beyond the lexicon to grammar (Langacker, 1987), and morphological and syntactic structures alike are viewed as meaningful. I subscribe to the major tenets of CG specifically the ones represented here. They serve as the primary conceptual framework for the present study.

Thus, within the framework of CG, the symbolic nature of language extends beyond the lexicon to morphology and syntax. The continuum that lexicon, morphology and syntax form reduces fully to the assemblies of *symbolic structures* (ibid: 12). Consequently, morphological and syntactic structures alike are "inherently symbolic, above and beyond the symbolic relations embodied in the lexical items they employed" (ibid: 12).

The key to the Cognitive Grammar approach lies in the notion of the symbolic "unit". Symbolic units associate a phonological representation with a semantic representation and vary in their degree of schematicity: Most lexical words, such as *tree, house*, and *run*, are richly specified, both semantically and phonologically. In contrast, most so-called function words, such as *the, a,* and *of* while phonologically fully specified, are semantically schematic. A syntactic construction, such as transitive or dative, is

specified at a higher level of schematicity. So, a symbolic unit comprises not only the sound-meaning associations of lexical items but also the formal and semantic aspects of constructional schemas.

Both idioms and constructions can be regarded as symbolic units, which associate a phonological representation with a semantic representation. Constructions usually are specified at a high level of schematicity and are able to sanction an open set of expressions. Idioms generally need to be specified at a lower level of schematicity. In the limiting case, a fully specified phonological form is associated with a fixed, conventional meaning. (It is rarely the case that an idiom is unanalyzable, whether in its formal or semantic aspects, so even idioms can be brought under more general schemas).

The difference between idioms and constructions turns out to be a gradient distinction having to do essentially with the schematicity at which a unit is specified. So, knowledge of (the syntax of) a language amounts to knowledge of a vast inventory of construction-idioms (or idiom-constructions).

Lexicon and syntax differ merely with respect to the schematicity and internal complexity of the semantic and phonological units that are associated. The notion of symbolic unit is easily able to accommodate idioms – that is, multiword expressions that speakers have learned as conventional associations of a phonological form with a semantic representation. Once it is recognized that constructions too are kinds of idioms (at a schematic level), it becomes possible to give idioms their proper place. Rather than being peripheral to the "core" of a language, it becomes possible to argue that idioms are the core. It been argued by both Construction Grammar and Cognitive Grammar approaches that all grammatical knowledge of a speaker can be represented by means of

constructions, the majority of which happen to be idiomatic in various senses of this concept. Thus, a person's knowledge of a language consists, precisely, in knowledge of idioms, that is, conventionalized form-meaning relations, at varying levels of generality. Everything turns out to be idiomatic, to a greater or lesser extent. A construction's usage range may not be fully predictable; constructions, in other words, display varying degrees of idiomaticity.

2.4.1 Lexicon vs. syntax

How does the variety of phenomena under the rubric of idiomatic fit into the theory of grammar? As was shown in Section 2.1 generative models compartmentalize lexicon and syntax. The lexicon lists the words of the language and states their individual properties – their pronunciation, their meaning, the lexical category to which they belong. The lexicon is the domain of idiosyncratic aspects that may be relevant -- such as their membership in an inflectional class.

The syntax deals with the combination of words into larger units. The syntax is regarded as the domain of the regular and the predictable. It is here where generalizations can be sought and rules can be formulated. Whereas speakers of a language manifestly have to learn the words of their language, they do not have to learn the sentences. The sentences can be generated by the application of general rules for the combination of the words. In this approach, syntax is the task of discovering and formulating these general rules.

A linguistic theory such as Generative Grammar, which compartmentalizes syntax and the lexicon, has no way of accommodating ready-made expressions of the nature described above. As we have seen from the earlier examples, some ready-made

expressions with regular syntax have to be listed in the lexicon as a unit, due to their non-compositional semantics. In the case of constructions, (transitive, dative, resultative, etc.) we have to make reference to the semantics of specific verbs that appear in the specific constructions and to the semantic role of their subject nominals. This upsets the neat compartmentalization of lexicon and syntax.

This means the distinction between lexicon and syntax can no longer be maintained. The syntax makes reference to lexical peculiarities, while the lexicon makes references to syntactic generalities. Neither is it the case that idioms are always characterized by idiosyncrasies, whether syntactic or semantic.

We have to turn to a different theory of grammar to accommodate the idiomatic expressions in a language. These kinds of problems do not arise in Cognitive Grammar. In Cognitive Grammar, knowledge of the idiomatic expression is part of the knowledge of English. Cognitive Grammar views each of these approaches as based on problematic assumptions.

The first one rests on the assumption that expressions can be cleanly divided into those that exhibit semantic compositionality and those that do not. The compositionality principle rests on the assumption that each component of an expression can be assigned a fixed, determinate and context-independent meaning.

The second rests on the assumption that syntactic rules are indeed general in their application and operate independently of idiomatic considerations.

Both of these assumptions turn out to be false. Once this is recognized, we must question whether it is indeed useful, or even possible, to distinguish the idiomatic from the non-idiomatic.

2.5 Access, Processing and Choice

There are a number of studies in the literature that have investigated the role of literal processing during the comprehension of idiomatic phrases.

Swinney and Cutler (1979) argued that idioms are stored as single entries in the mental lexicon, and are retrieved in the same manner as any other lexical item. They refer to this view as the *lexical representation hypothesis*. Thus, according to this view, the comprehension of an idiom does not require that a compositional analysis be undertaken. In this model, the figurative meaning of an idiom can be retrieved directly (and holistically) from the lexicon. Swinney and Cutler assumed that the retrieval of an idiom's figurative meaning occurs in parallel with the computation of its literal meaning. Figurative processing, however, will tend to conclude more quickly than literal processing because the direct access of the figurative meaning is simpler than the compositional analysis required to derive a literal interpretation.

Peterson and Burgess (1993) developed three new models that specify the relationship between syntactic and semantic processing during the comprehension of formulaic phrases. They refer to these models as the *syntactic dominance,* the *semantic dominance,* and *syntactic-semantic autonomy* models respectively. The three models make unique predictions with respect to the nature of syntactic/semantic dependencies in on-line comprehension of formulaic language. The syntactic dominance and the syntactic-semantic autonomy models propose that a full-description of the syntactic structure of an idiomatic phrase is necessarily computed in a figuratively biasing context. On the other hand, the semantic dominance model suggests that syntactic analysis is terminated during the course of processing of an idiomatic phrase in such context.

Another issue central to these three models is whether the full literal meaning of an idiom is computed obligatorily. The syntactic dominance model claims that it is, whereas the other models suggest that termination of literal analysis is possible. Peterson and Burgess experimental work yielded mixed results. They found dissociation between syntactic and semantic analyses during on-line processing of formulaic expressions. Their results showed that the processing of idiomatic and literal strings were largely indistinguishable in terms of the syntactic representations that were derived. The literal meaning of an idiomatic phrase was computed only in literally biasing contexts. When the contexts biased an idiomatic interpretation of the phrase, literal analysis appeared to be terminated prior to the final word of the idiom. A conceptual priming effect was found for literal sentences only. Taken together, these results suggested that, by the final word of an idiomatic phrase, subjects continued to monitor the syntactic structure of the idiom, but do not compute a corresponding literal interpretation of the phrase. Peterson and Burgess interpret their experimental results in favor of the syntactic-semantic autonomy model of processing, which is compatible with modular views of parsing. This view stipulates that the parser is insensitive to semantic and pragmatic contingencies (Frazier, Clifton, and Randall, 1983).

The basis for the current project is the study by Erman and Warren (2000) that seeks to discover the nature of formulaic language in spoken and written English discourse and its interaction with a non-formulaic language. Two mechanisms for language storage, processing, production and use are made use of in this study – "the idiom principle" and "the open choice principle" (Sinclair, 1991). According to Sinclair (1991), the entire language system is organized according to two principles: the idiom

principle, which covers multiword, preconstructed combinations that are processed as unified wholes or chunks without an analysis of the component parts and the open choice principle, which includes word-for-word combinations that are processed analytically one component at a time. The processing of a unified preconstructed chunk of various length represents a single choice for the speaker and the hearer and therefore is considered to be where every word represents a choice. The alternation and the relationship between the two principles is the focus of the Erman and Warren's investigation.

It has been argued that the idiom and the open choice principles represent one and the same type of lexical access (Bybee and Torress, 2009). However, the majority of researchers on formulaic language distinguish the two principles as different processing types: The idiom principle closely corresponds to holistic processing and the open choice principle closely corresponds to analytic processing. The idiom principle engenders holistic processing because a unified pre-constructed chunk or multiword string of various length represents one choice for the speaker and the hearer. Whereas the open choice principle operates on a one -word –at- a- time bases where each word represents a separate entry in a lexicon and therefore a separate choice. Formulaic language is thought to be processed faster than non formulaic language based on the fact that the reduced number of choices must be made by the speaker in composing formulaic language. Thus, it reduces cognitive costs necessary in encoding and decoding a message.

2.6 Word Order and Typology

The research on word order variations evolves around three major questions: What is the basic order in language? When there are several possible order patterns in a language, what is the communicative function of one, rather than another, order? What

historical reanalysis gives rise to observed order patterns (Payne, 1992)? The initial question of identifying a basic order of subject and object relative to the verb started the tradition of language typology as exemplified by Joseph Greenberg (1963). The tradition has been continued by numerous scholars, notably including Lehmann (1973), Vennemann and Harlow (1977), Hawkins (1983), Nichols (1986), and Dryer (1988). Various definitions of "basic" ignited controversy among scholars, leading some to differentiate the typological division between those languages in which the main clause word order primarily correlates with pragmatic factors and those in which order primarily correlates with grammatical relations or other syntactic factors (Thompson, 1978; Payne, 1990, 1992.) Word order variation is a complex phenomenon and needs to be examined from syntactic, cognitive and pragmatic viewpoints. Payne (1992:2) states that "explanatory factors behind word order variation are to be found in studies of how the mind grammaticizes forms, processes information, and speech act theory considerations of speakers' attempts to get their hearers to build one, rather than another, mental representation of incoming information." A combination of these three domains will contribute to determining the word order in a language and its flexibility.

It is generally assumed that for a majority of the world's languages, one can identify a "basic" order of subject and object relative to the verb. A common diagnostic of basic order is statistical frequency (Dryer, 1983). Whichever order appears the most often might be considered basic. However, languages differ greatly in terms of the flexibility of the word order within a language that speakers render grammatical. The continuum goes anywhere from an entirely fixed order of constituents to a situation where speakers agree that all logically possible constituent orders are grammatical: SOV,

SVO, VSO, VOS, OSV, and VSO. It seems plausible to assume that there is a correlation between the type of language and the mechanisms that speakers employ in the production of a particular type of language. The alternation between the use of word-for-word combinations and the use of preconstructed multiword combinations might vary in typologically different languages. Therefore, the use of the open choice principle vs. the idiom principle might vary in typologically different languages. To test these assumptions, I put forward the following hypothesis:

1) A language with a free word order will use idiom principle in the construction of sentences just as much as a language with a fixed word order. A free word order language will use as many prefabs as a fixed word order language.

The hypothesis can be reformulated in the following way:

2) A language with a free word order will use open-choice principle to a greater extent than a language with a fixed word order. Thus, a free word order language will have a lower percentage of prefabs than a fixed word order language.

The comparison of test results of typologically different languages will reveal the universality of various mechanisms in the production of language.

2.7 Summary of Chapter 2

Chapter 2 is a review of the treatment of idiomaticity in various linguistic theories. This survey follows a historic chronology starting from Generative Grammar, which dominated linguistic thought most of the 20th century, to subsequent usage-based approaches, which include a variety of construction grammar theories as well as cognitive grammar framework. While the model of generative grammar views

idiomaticity as problematic, exceptional and non-significant, functional approaches embrace it as central, unexceptional and paramount for the study of natural language; indeed, the "core" part of human language. Idiomaticity can be described in many different ways, and several typologies have been proposed; semantics, pragmatics, psycholinguistics, as well as syntax have alternately been the focus of various typological classifications.

Idiomaticity has been equated with formulicity in the field of phraseology that originally investigated the phenomenon. Formulaic language is contrasted with non-formulaic where two represent distinct mechanisms of linguistic storage and processing. Formulaic language is viewed as organized according to the idiom principle, a type of language that is stored and processed holistically, i.e. without recourse to its internal configurations. Non-formulaic language is organized according to the open-choice principle, a language that is stored and processed analytically. Formulaicity can be best represented by a continuum where an internal structure can be more or less fixed, from completely frozen items that do not undergo any syntactic transformations all the way to items that are formulaic but syntactically flexible. Idioms by definition are non-compositional. However, idiomatic expressions vary greatly in their analyzability where some instances are quite transparent and others are obscure and cannot be figured out without reference to their etymological origins. The interaction between the two mechanisms - formulaic and non-formulaic -- has been the focus of some studies such as Erman and Warren (2000), Pawley and Syder (1989), Wray (2000), Jekendoff (1995), Melcuk (1995), and Foster (2001). These and other studies analyze various types of texts of the English language and provide reliable quantitative data on the proportion of

formulaic sequences in spoken and written discourse. The phenomenon of formulaic language, however, has not been studied cross-linguistically.

The current study proposes to investigate the interaction of two mechanisms – formulaic and non-formulaic -- in a free word order language, which contrasts with the English language in myriad ways. Russian, a conservative Indo-European system, has been chosen for several reasons. Russian provides the necessary contrast in the flexibility of the word order organization and the morphological complexity. It is traditionally viewed as a free word order language in typological research. As a native speaker, I can use my judgment in the analysis of Russian prefabs, which I propose to analyze in a variety of spoken and written genres.

The analysis of prefabs in Russian follows a method of Erman and Warren (2000) and seeks to quantify the proportion of formulaic and nonformulaic material in a free word order language. In addition, processing of prefabs represents a reduced cognitive load in comparison to processing a non-formulaic material as it represents a significantly reduced number of choices for the language user. The method of analysis accounts for the number of overall choices as well as specifically lexical choices made in the production of texts. Issues of access, processing and choice are discussed in the light of the analyses of language type, discourse medium, genres and speakers.

Chapter 3

Data and Method

3.0 Introduction

One of the goals of the study is to apply the mode of analysis developed by Erman and Warren (2000) for the analysis of the English prefabs in spoken and written texts to a typologically different language. The premise of this approach is that language learners have a number of more or less preconstructed phrases available to them, and that the production of texts involves the alternation of prefabricated and non-prefabricated language. The method was originally designed to discover the impact that this alternation has on the structure of a text as a whole. Since prefabricated expressions – multiword composites – do not always converge with grammatical phrases, the traditional structural analysis is not always applicable to their identification, classification, and analysis. However, we must ask what the status of these prefabs is in the adult language. First, the interest of the study is to determine in what ways the various categories of formulaic expressions interact with each other and with the language that is novel or analytic. The second goal is to determine how the method for the analysis of prefabs developed by Erman and Warren (2000) can be applied cross-linguistically. Specifically, can his method be used in the analysis of prefabs in a language of a different morph-syntactic type with a less restrictive word order than English. If the method is applicable to a different linguistic system, then results can be compared and analyzed, revealing how prefabricated language and language type interact. Any correlation between the number and the variety of formulaic expressions and a particular language type can contribute to the construction of the model of language that is universal in a sense that is common to

human experience despite cultural and cross-linguistic variation. Thus, if results of the analysis hold cross-linguistically, this puts us in the position to seek explanation in cognitive and interactive domains common to human experience rather than within the language-specific structural patterns of a language.

Erman and Warren (2000) classify all prefabs into four categories – *lexical*, *grammatical*, *pragmatic* and *reducible*. The classification allows identification of the proportion, distribution, and length of prefabricated material in a text. Lexical prefabs refer to the extra-linguistic entities such as things, events, states, and situations. For the reason that lexical prefabs often have a real world referent they are easier to identify and to classify. On the other hand, functional prefabs, which include grammatical and pragmatic prefabs, are intra-linguistic entities and serve various functions depending on their nature and use.

The existence of various types of prefabs reflects different reasons for conventionalization of phrases. This classification was modified for the present study to accommodate the analysis of prefabs in the Russian language. The present study seeks to utilize Erman and Warren (2000) classification as much as possible, so results can be compared. Some modifications were necessitated by the differences that lie in the structure of the two languages. My account of the formulaic sequences in Russian involves an analysis of three independent categories of prefabs: lexical, grammatical, and pragmatic.

3.1 Method

3.1.1 Considerations on the choice of typology

Several typologies of formulaic sequences have been considered for the present study, including Becker (1975), Bolinger (1976), Hatch et al. (1979), Coulmas (1979, 1994), Yorio (1980), Nattinger and DeCarrico (1992), Lattey (1986), Van Lancker (1987), Moon (1992, 1998), Howarth (1998), Wray (2000, 2008), and Fillmore et al. (1988). Some researchers characterize formulaic sequences based on their form, while others focus on their functions. Frequently, the two parameters are conflated.

For example, Nattinger and Decarrico (1992) propose a classification of formulaic sequences without teasing apart considerations of form and function to include three categories of idioms. Becker (1975) draws heavily upon Nattinger and Decarrico's classification but expands it to include six categories in his typology: polywords, phrasal constraints, meta-messages, sentence builders, situational utterances, and verbatim texts. Six-way characterization is more detailed than Nattinger and Decarrico (1992). However, it too fails to consider issues pertaining to form and function, or to put formulaic sequences on a continuum from fixed to novel.

Another typology that was considered for the study was by Fillmore et al. (1988), who provided a four-way characterization of formulaic language. First, from the semantic point of view, these researchers characterized formulaic sequences as encoding idioms and decoding idioms (following Makkai, 1972). Encoding idioms such as *answer the door, wide awake, bright red* can be figured out by a speaker on the first hearing, but the speaker may not know this is a preferred, conventional way of expressing a particular

idea in a language. Decoding idioms, on the other hand, cannot be understood by the hearer but must be explicitly taught, i.e., *kick the bucket, pull a fast one.*

It is reasonable to view idiomaticity not only as a semantic but as a syntactic phenomenon because units of formulaic language are larger than a word and have an internal structure. Often idioms can be entirely fixed or deviate from regular syntactic patterns.

The second distinction Fillmore et al. (ibid: 504) make classifies idioms as grammatical and extragrammatical based on syntactic patterns. Grammatical idioms follow the regular rules of syntax but are idiomatic in other ways. Extragrammatical idioms have anamoulous structure that deviates from regular syntactic patterns, i.e., *first off, sight unseen, all of a sudden, by and large, so far so good.*

The third distinction in Fillmore's et al. (ibid: 505) classification is based on the schematicity continuum. Substantive idioms represent one end of the continuum, while formal idioms (i.e. same as schematic idioms in Croft, 2001) represent the other. Substantive or lexically filled idioms are fully specified. All of the idioms used as examples so far are substantive idioms. Formal idioms, on the other hand, are schematic patterns that have semantic or pragmatic value and are not predictable by their form alone. For this reason, formal idioms are also referred to as schematic idioms. Substantive idioms are thought of as "large words" and are represented and stored in the lexicon in the same manner as single words. Formal or schematic idioms, on the other hand, vary widely on the scale of schematicity, some with just a few schematic elements and others, completely schematic, without any lexically specified items, *the X-er, the Y-er*; *NP VP NP*. Completely schematic structures include such constructions as transitive, ditransitive,

and resultative, etc. Fully schematic patterns, known as constructional schemas, are stored and retrieved whole from memory.

The forth category of idioms that Fillmore et al. (ibid: 506) postulate are the ones "with or without pragmatic point." This type of idiom exists for specific pragmatic or rhetorical purposes. For example, *Him be a doctor?* or *Didn't you like the salad?* do more than ask a yes/no question. While many substantive idioms are used for specific pragmatic purpose (*Good morning; How are you?; Once upon a time*), schematic idioms are usually free of pragmatic commitments (*the X-er, the Y-er.*) While being a detailed characterization of idioms, this typology was ultimately rejected for the present study because it does not provide a method for quantifying formulaic language; i.e., the comparison across texts and languages becomes impossible without a quantitative measure of formulaic sequences.

Erman and Warren (2000) propose a typology based on four categories: lexical, grammatical, pragmatic, and reducibles. The four-way classification may seem limited in light of the plethora of forms and functions that formulaic language can take. A choice of typology, however, ultimately depends on goals of a study, research questions, and methodology used. Erman and Warren's four-way characterization is sufficient for the purpose of establishing the number of formulaic sequences in a text, for examining the distribution of prefab types across texts and genres, and for measuring length of prefabs and non-prefabs.

Erman and Warren's typology is not without problems. At times it is difficult to determine whether a prefab is grammatical or lexical and sometimes whether it is grammatical or pragmatic. The process of grammaticization affects lexical prefabs to the

point where they lose their semantic content and acquire more functional properties. Another difficulty lies in the fact that one and the same prefab can have multiple functions and can be a part of multiple groups. Erman and Warren overcome these issues by consistently assigning prefabs with debatable nature to a particular category.

Erman and Warren's typology was adopted for the present study for several reasons. First, one of the goals of the current study is to establish the proportion of prefabricated material in a language typologically different from English. The method used for this purpose in the study of English has never been applied cross-linguistically before. Less detailed rather than more detailed typology is more useful for the task.

Second, to determine how to apply Erman and Warren's method to a typologically different language and ensure the comparability of the results between English and Russian data, the same typology must be used. While Erman and Warren's typology is not the only one possible, it serves the purposes of the current study to document proportion, distribution and length of formulaic sequences in a free-word order language and makes comparability of the results between texts and languages possible.

3.1.2 Structural principles: open choice vs. idiom continuum

There is no established tradition in psycholinguistics of studying how humans process units larger than single words but smaller than sentences (see Tanenhaus, 1988, for a review.) Two mechanisms have been proposed for processing multiword sequences — *holistic*, where chunks of language are processed as a single unit without recourse to their internal composition, and *analytic*, which follows standard combinatorial rules of traditional generative grammar. These two types of processing correspond to what Sinclair (1997) refers to as "idiom principle" and "open choice principle," respectively.

Formulaic language is processed holistically either due to some type of semantic or syntactic irregularity or to the level of entrenchment of regular multiword sequences that reached the status of a "unit" (Langacker, 1987). In case of semantic or syntactic irregularity, the expressions are not analyzable into segments, as often is the case with certain idioms. Analyzability of non-compositional phrases is low or does not exist at all. Thus, a non-analyzable unit can be represented and processed only holistically. In addition, high frequency, regular multiword sequences may be compositional but lost their analyzability due to their high frequency of use. They also tend to have a holistic representation and are processed as one unit.

On the other hand, the principle of open choice corresponds with analytic processing of novel expressions. Due to their novelty, the expressions are compositional and analyzable into their constituent parts. The two types of processing represent two ends of a continuum. Between the prototypical instances of utterances processed according to the idiom and the open-choice principles lies a range of expressions that can be more or less formulaic, that allow lexical substitutions and syntactic manipulations. Thus, fixed and formulaic phrases may have variable elements that represent a speaker's choice of a lexical item or grammatical element. For example, the phrase *He went to some seminar/lecture/meeting/etc*. contains a restricted variability slot, which can be filled with various but related items. Thus, a formulaic construction that represents a unified structure is reflective of an idiom principle, while variable elements within a construction are chosen according to the open-choice principle. Prefabs with open or restricted variability slots represent some flexibility and choice for the language user. The analytic-holistic continuum reflects the graded nature inherent in the phenomenon of

formulaicity. The interaction of the two types of processing is the focus of the recent research program in formulaic language (Wray 2002, 2008; Peterson and Burgress, 1993; Erman and Warren, 2000).

3.1.2.1 Prefabs vs. schemas

Prefabs by definition represent units of various length that have at least two words that are lexically filled: *answer the door, bright red, wide awake*, etc. Prefabs vary in length. Most typical prefabs contain two to four words. Some, however, can be up to 14 or 16 words in length.

Prefabs can have open slots and restricted variability slots, which are not considered parts of prefabs; i.e., they were not counted in the prefab analysis. Open slots as well as restricted variability slots can be seen as "fuzzy" at the boundary between prefabricated language, on one hand, and novel language on the other. Restricted variability slots, for instance, represent a choice, but the choice is limited or defined by a prefab itself, which serves as a frame in a sense of Fillmore et al. (1988). Thus, restricted variability slots are not typical open-choice slots; at the same time they are not idiom slots either because some variability is possible. They represent the graded nature of the open-choice vs. idiom principles and the processing mechanisms associated with the continuum. Erman and Warren (ibid: 32) note that prefabs are not identical to the constructions in Construction Grammar analysis. The reason is that constructions such as transitive, ditransitive, relative, etc. are not lexically specified. For example, transitive construction NP+ VP + NP; *I ate a pie*; ditransitive construction NP+VP+NP+NP; *I read my son a book*, etc. are constructional schemas that do not have obligatory lexical material. Rather, they are templates that can sanction the formation of new utterances.

Constructions or constructional schemas such as these arise as generalizations over a great number of instances of such constructions in the actual use of the language. Langacker (2008) describes constructional schemas as imminent and indissociable from the instances of the use of constructions. Fully schematic constructions do not represent qualitatively different material from prefabs, but rather they are more schematic than lexically specified prefabs. As part of the design of the current study of prefabs, fully schematic constructions are excluded as the end point on the generality-specificity continuum. The exclusion of fully schematic constructions follows the original choice in the methodology of the study of the English prefabs (Erman and Warren 2000). In contrast to the lexically unspecified constructions, prefabs by definition must have lexical material; at least two slots in a prefab must be lexically specified, where the choice of one word determines or at least restricts the choice of another, usually adjacent, word. Some constructions, however, do have lexically specified elements – *let alone* construction, *if it is good for X, it is good for me*, etc. Such types of constructions meet the definition of a prefab and are included in the prefab analysis. It should be clear that the distinction drawn between fully schematic constructions and lexically specified prefabs in this method is an artificial one; it is used to limit the scope of the study. Fully schematic constructions represent an end of the continuum of schematicity, whereas prefabs are less schematic examples of the same continuum.

The presence of open slots and, especially, the open slots with restricted variability indicate that prefabs have fuzzy boundaries and in this way can potentially be viewed as constructions. Prefabs represent specific lexical, grammatical and pragmatic material over which generalizations can be made and out of which constructional

schemas arise. Clearly, we are dealing with a graded phenomenon that is best represented by a continuum with prefabs being more specific units on a continuum and constructional schemas being more schematic. Specific multiword units with high collocational frequency give rise to patterns that syntacticize and form constructions (Givon, 1978). Thus, the relationship between prefabs and constructions is one of schematization; prefabs are instances of use that can serve as material for generalizations that form constructions.

3.2. A Prefab Analysis in English and Russian

This section describes conventions used to analyze English prefabs. The same conventions were used for the analysis of the Russian prefabs in the current study. English examples are adapted from Erman and Warren, and the Russian examples are used from the Russian data.

In this method, each word in a text represents *a slot.* A slot can be filled by a word chosen according to the open-choice principle or a word that is part of a prefab. Dashes [–] are used to replace words that are chosen according to the open-choice principle. The alternation between prefabs and non-prefabricated elements are demonstrated in (1). The beginning and the end of a prefab are marked by a slash:

(1) /you know/ /I went to some *seminars*/

--- --- --- --- /a waste of time/ /are they/

The first step in the analysis is to establish the number, type, and proportion of prefabs in a text. In example (1) there are 17 slots and four prefabs – two lexical, one pragmatic, and one grammatical. The breakdown of the prefabs is the following:

Pragmatic – you know

Lexical – I went to *seminars*; a waste of time
Grammatical – are they

In this example, the second prefab -- /I went to some *seminars*/-- contains an *open slot*, which is filled by a lexical item "seminars". An open slot represents a slot that must be filled by lexical material for the prefab to be complete but can be filled by a practically unlimited number of words. This particular example, however, represents what is known as a "restricted variability slot"-- a slot that can be filled with a word of a particular category: *seminars, lectures, meetings, workshops, etc.* Such words are not removed but are italicized and reduced in size. These words are ignored when the number of slots filled by parts of prefabs is counted. Thus, in example (1), out of 17 slots, only 12 are filled with portions of prefabs. However, when the analysis of choices is performed, i.e. the number of choices made by a speaker, these words are not ignored but are counted as a choice.

Some open slots can be filled by prefabs themselves. To indicate this, square brackets are used and the size of the prefab is not reduced, as in example (2).

(2) / lead [that sort of] life /

/ [the answer to this question] depends on [one's answer to question…] /

Such embedding of prefabs is common and leads to formation of structures of various levels of complexity.

In addition, prefabs can be often extendible. Extensions, which are not obligatory, are put within parenthesis, as indicated in example (3):

(3) (quite) all right

due (mainly) to sth.

3.3 Analysis of Choices in English and Russian

The second analysis in Emran and Warren's methodology is to establish the number of choices made in producing a text. The choice represents a cognitive effort made by a language user in encoding and decoding a message. Choosing and retrieving each word separately represents a greater effort than choosing a preconstructed, ready-made sequence of four words such as *How are you doing?*, which is routinely used as a greeting formula. A greeting formula such as this represents a single multiword retrieval from a mental lexicon of a speaker. Because of the presence of prefabs in any text, whether written or spoken, produced by any language user of any genre, the number of retrievals is always less than the number of words in a text. According to the convention, underlinings are used to mark a choice, i.e., each choice is underlined by a solid line as in example (4). Plus signs (+) as in "to the best of + my + knowledge" indicate that this string involves two choices (the prefab and the variable determiner) and not three.

(4) To the best of + my + knowledge, there is no record of a society which has used literacy for + the profane and imaginative + purposes and which has + not + produced books dealing with sexual topics;

Thus, the difference between counting words and counting choices is significant. Not every word represents a choice. The use of prefabs in any given text means reducing the number of choices a language user has to make since a prefab by definition is a combination of at least two orthographic words. Thus, example (4) illustrates that the utterance contains 33 slots, and only 23 choices were made by the speaker. The number of choices is indicative of the processing effort involved in producing a text. The comparison between text types, genres and speakers can be made with respect to the

numbers of choices made in producing the text. In this method, one word can never represent more than one choice; therefore, the choice of inflections such as 3d p. sing. *-s* or progressive –ing. are ignored. In other words, examples such as *walks, walked, has/had walked, will walk* represent one choice regardless of whether an auxiliary is present or not.

3.3.1 Analysis of lexical choices in English and Russian

The third step in the analysis is to establish the number of lexical choices made, in addition to counting all choices made in producing a text. Lexical choices represent greater cognitive effort (with the possible exception of restricted variability) since choosing an item from a restricted set is less demanding than selecting one from an indefinitely large set.

In example (5), lexical choices are marked by the bold print:

(5) To the best of + my + **knowledge**, there is no **record of** a **society** which **has used literacy** for + the **profane** and **imaginative** + **purposes** and which **has** + not + **produced books dealing with sexual topics;**

Thus, example (5) contains 33 slots, 20 of which are filled by prefabs; and 23 choices, 13 of which are lexical. It also has six prefabs with four lexical prefabs, one grammatical and one reducible.

The breakdown of the prefabs is as following and is explained in more detail in the following section:

Lexical

to the best of my/our etc. knowledge
a record of sth
for the *profane* purposes

books / articles etc. deal with *sexual* topics/matters etc.

Grammatical

there is (exist) NB reducible.

Reducible

has not.

This classification is not the only one possible and the estimated number of choices is not indisputable. However, this methodology is meant to be indicative of the processing effort involved in producing a text and to make a comparison of texts, genres, and languages possible.

3.4 Compositional Prefabs

Some prefabs can be combined to form even a larger single unit: *a little bit + more than, that one + over there, due to +the fact that, books + deal with some topic.* Thus, prefabs can combine with each other and/or other words to form a *composite prefab*. Combining prefabs can become complex with multiple embedding of prefabs and multiple open slots. Such nesting of prefabs is indicative of the complexity of the formed linguistic structure as the result of language use. Embedded prefabs are marked by square brackets and words in open slots are italicized as illustrated in Examples 6-8. It is common for a prefab to fill an open slot of another prefab.

(6) /[*the Prime Minister*] and *Mr. Lloyd* had a (whisprered) conversation/

/*they* nodded at [each other]/

(7) /the preparations for [*launching their rockets*]/

(8) /the average of [forty miles [an hour]]/

Prefabs can overlap or occur successively as Examples 9 and 10 respectively illustrate. An overlap in English prefabs occurs exclusively with reducibles, where the end of one prefab is simultaneously the beginning of another.

(9) /*Hart* [you[‘VE] got to] stand up to/ /haven’t you/

(10) /I gather/ /[*you’ve* been at it/ /for *nine* years/

/By golly/ /[that’s] true/

As mentioned earlier, it is not unusual for an open slot of a prefab to be filled with another prefab. At the same time, a prefab used to fill an open spot can itself have an open slot.

This is a situation of a *double open* illustrated in example 11:

(11) /[the first] hint of [*the Chancellor bowing to* [*public opinion*]]/

double open

These examples illustrate how the complexity involved in the alternation of prefabricated and non-prefabricated string. Prefabs and non-prefabs are typically interspersed as in Example (12):

(12) /Just as/ /*she* was wondering what to do (next)/ /*she* came upon the *Three Bears’ little house.* /-- -- -- -- -- -- -- / how nice it would be to *have someone to say*,/ / ‘Come in,/ /my dear/. / Sit down/ -- /have some breakfast.’ -- /no one/ -- -- -- -- / in answer to *her gentle knock*/ -- /*she* ran round/ -- /the window peeped through/ -- -- -- -- -- -- --

What is interesting in this and other examples in the analysis of English data is that non-prefabricated newly constructed language is not a dominant way of using language. This is especially evident in a spoken medium. In spoken language, in fact, the

longest example of a non-prefabricated string contains 11 slots, but this is rather exceptional. The great majority of nonprefabricated strings consists of one or two members. Nonprefabricated strings tend to be somewhat longer in written than in spoken language overall. The longest nonprefabricated string in written language in the Erman and Warren's analysis is 14 slots. The most important difference between spoken and written language that Erman and Warren's method demonstrates is that nonprefabricated strings are longer in written language, which is expected if we consider the differences involved in producing the two types.

High frequency of use of various formulaic sequences led Wray (2000) to conclude that the use of formulaic language in everyday life is a default way of organizing and constructing language. She argues that reliance on formulaic language is a preferred strategy of dealing with the expected and predictable, which abounds in everyday experiences. The recurrent ways of dealing with recurrent situations becomes conventionalized in repeated linguistic forms (ibid: 74). When things are unexpected, then the language user falls back on the use of newly created, analytic language.

"Formulaicity characterizes the normal approach to processing, with analyticity on hand to pick up any difficulties, such as can be caused by a speaker's thick accent or non-native grammar, background noise, dysfluency, poetry, word games, and so on. Our baseline strategy in everyday language processing, both production and comprehension, 'relies not on the *potential for the unexpected* in a given utterance but upon the *statistical likelihood of the expected*" (Wray 2000).

In fact, Erman and Warren's analysis shows that the longer the non-prefabricated string, the rarer it is. An existence of open slots and slots with restricted variability

tangibly demonstrates that the distinction between novel and formulaic language is not categorical. Formulaic sequences can have open slots, which present an opportunity for a novelty in an utterance to be manifested. This is what makes prefabs flexible and adaptable to incorporating new meanings or new situations. On the other hand, restricted variability is not a complete open choice but a choice within a category. Restricted variability allows some flexibility within a category, which limits the choices. Example 13 demonstrates restricted variability in initial NP position.

(13) *book, novels, articles, speeches, talks* + deal with some topics

The variety of formulaic sequences represents a continuum from completely invariable, "frozen" expressions to prefabs that contain open slots with limited choices (restricted variability), and further to expressions with open slots, which have no limitations and can be filled by practically any word. This continuum reflects the degree of flexibility that is built into formulaic sequences to make them adjustable to a variety of contexts.

This method demonstrates the alternation of prefabricated and non-prefabricated language in production of texts. It gives statistical results of how the strategies interact, revealing the prevailing tendency for the use of formulaic language in both spoken and written data.

The multiple uses of a variety of formulaic sequences and the functions they serve will be explored in the course of the study.

3.5 Types of Prefabs

Erman and Warren's (2000) typology includes a four-way characterization of prefabs: lexical, grammatical, pragmatic and reducibles. Different reasons exist for

conventionalization of different types of prefabs. Considerations of meaning, form and function are reflected in this classification. In the next sections I describe each type in more details.

3.5.1 Lexical prefabs

Erman and Warren define lexical prefabs as "semantic units in that the y have reference and denote entities, properties, states, events, and situations of different kinds…" The emphasis is made on making certain that the lexical prefab does represent some extra-linguistic entity or phenomena, since the precision of the notional characterizations of prefabs is difficult to achieve.

The situations that can typically be defined by lexical prefabs:

- properties and states – *out of date, be of help to somebody, different from something, have got something*
- situations and events – *find one's way out of something, run off, make sure*
- places and positions –*here and there, at headquarters, to the right, in industry, on paper.*
- entities – *sketch pad, modern furniture, phone call, permanent job, subject matter*
- period or point of time – *at the time, by then, in the end, the eighteenth century*

This category can be used in the analysis of another language unmodified due to the fact that lexical prefabs often represent a real-world entity or phenomena. This category was adopted to the analysis of Russian prefabs without modifications.

Some lexical prefabs can grammaticise over time and acquire functional status. For example such English prefabs as *in reply to, of cause, on these grounds* have a

functional rather than a lexical role. These instances can be difficult to place in a lexical category, however, it was done by Erman and Warren for the purpose of consistency. Other prefabs can acquire a pragmatic function: *that's true, it's all my fault, don't bother, is that all? that's a good idea, what is it all about?* The distinction, therefore, between lexical and functional prefabs can be fluid, as often is the case with grammaticising units. In this study, the decision was made to keep questionable cases like this in a lexical category for consistency as well.

Despite the fact that prefabs do not always coincide with traditional grammatical categories, it is possible in some cases to characterize lexical prefabs syntactically in terms of clause of phrase structure, for example:

- lexical noun-phrase prefabs – *a waste of time, present state of our knowledge, rule(s) of something*
- lexical verb-phrase prefabs – *fail to do something, be in touch with, get the hang of something*
- lexical preposition-phrase prefabs – *for some reason, to the naked eye, on a clear night*
- lexical adjective-phrase prefabs – *ignorant of something, able to do something, enough of something, suitable for something*
- lexical adverbial-phrase prefabs – *once again, so far, all over the place*

This classification is useful and can add to our understanding of formulaic sequences as long as they structurally overlap with the traditional phrasal categories. Because this often is not the case, additional criteria for identification and analysis of prefabs are necessary.

3.5.2 Grammatical prefabs

In contrast with lexical prefabs, grammatical prefabs are what Erman and Warren term "intralinguistic, text-forming" items rather than referring units. A grammatical prefab will "quantify, specify, modify the reference or meaning of nouns, verbs, adjectives, or adverbs, in a general manner, or they will serve as their substitutes or as links between propositional or referring items" (ibid: 57). English grammatical prefabs tend to be shorter and not as variable as lexical prefabs. Some grammatical prefabs can be extended, inflectionally modified, and some allow restricted choice of members. The majority of grammatical prefabs, however, are less variable than lexical prefabs and often quite frozen.

The following are the English grammatical function labels used by Erman and Warren:

- quantifiers -- a little, a number of
- links -- for example, neither…nor..
- determiners – some of, that sort of ,
- proforms -- and so on
- tense – be going to
- modals – be supposed to, have got to
- introductors – there is/are
- aspect – used to

The subgroup "aspect" was not used in the analysis of the Russian grammatical prefabs due to the fact that Russian marks aspectual distinctions morphologically. The subgroup "introductors" also was discarded because of the absence of this structure in Russian

altogether. However, several additional subgroups were added to accommodate such instances as case marking and subjunctive.

3.5.3 Pragmatic prefabs

Pragmatic prefabs form a functional category because they "do not directly partake in the propositional content of the utterance" (ibid: 57). They often occur outside of syntactic structure and thus differ from grammatical prefabs. Erman and Warren's three-way distinction to classify pragmatic prefabs includes textual, interpersonal, and metalinguistic. The same distinction was adopted for the analysis of the Russian prefabs. The English examples are:

1. Text monitors:

   discourse markers (*and then*)

   turn regulators (*well you know*)

   repair markers (*I mean*)

2. Social monitors:

   interactive (*wouldn't it*)

   feedback signals (*I see*)

   hesitations (*what's the word, you know, I mean*)

   responses (*yes, I think so; oh no, well yes, yes I see*)

   performatives (*do sit down, thank you, good luck, why don't you, good evening)*

3. Matalinguistic monitors:

   approximators (*and everything*)

   hedges (*sort of*)

epistemological signals (*I should think*)

attitudinal markers (*I must say, my dear*)

The peculiarity of pragmatic prefabs is that they can be multifunctional. *You know, you see, I mean* can be used as hesitation markers to stall for time and as repair markers to help speakers change wording or elicit audience involvement. Erman and Warren point out that one and the same marker can have more than one function, not only in different contexts but also in one and the same context. Thus, *you know, you see, I mean*, for instance, can serve as social monitors and text monitors simultaneously. Multifunctional nature of pragmatic prefabs is ubiquitous. Grammatical prefabs are not nearly as multifunctional as pragmatic prefabs. Pragmatic prefabs are substantially more frequent in conversation and thus serve multiple purposes that help meet physical, social, and cognitive demands of real time face-to-face interaction.

3.5.4 Reducibles

The last category used by Erman and Warren is *reducibles*, which include such English phenomena as contractions of pronoun and verb (*I'm, it's, they've, he'd, I'll*), auxiliaries and *not* (*don't, isn't, hasn't, can't, shouldn't, wouldn't*); auxiliary and auxiliary (*would've, should've*) and combination 'let' and 'us' (*let's*). The argument for the category *reducible* is based on some available evidence that such combinations are stored and retrieved from memory as ready-made items. These examples demonstrate that phonological reduction of frequently co-occurring elements leads to fusion that can be captured orthographically, as in English. Though we can expect phonological reduction and fusion to be present in any natural language, not every language formalizes this

process by means of orthography. This type of contraction is not present in Russian, for instance, and was not used as a category in the analysis of Russian prefabs.

Posing lexical, grammatical, and pragmatic prefabs can hardly be viewed as controversial. However, reducibles seem to be a language-specific, English category.

3.6 The Pilot Study

3.6.1 The hypotheses

It is generally assumed that for a majority of the world's languages, one can identify a basic order of subject and object relative to the verb. A common diagnostic of basic order is statistical frequency (Dryer, 1983.) Whichever order appears the most often might be considered basic. However, languages differ greatly in terms of the flexibility of the word order within a language that speakers render grammatical. The continuum goes anywhere from an entirely fixed order of constituents to a situation where speakers agree that all logically possible constituent orders are grammatical: SOV, SVO, VSO, VOS, OSV, and VSO. It seems plausible to assume there is a correlation between the type of language and the mechanisms speakers employ in the production of a particular type of language. The alternation between the use of word-for-word combinations and the use of preconstructed multiword combinations might vary in typologically different languages. Therefore, the utilization of the open-choice principle vs. the idiom principle might vary in typologically different languages. To test these assumptions I put forward the following three hypotheses:

Hypothesis 1: A language with a free word order will use idiom principle in the construction of sentences just as much as a language with a fixed word order. A free word order language will use as many prefabs as a fixed word order language.

Hypothesis 2: The idiom principle might take a different form in a free word order language than in a fixed word order language.

Hypothesis 3: There might be a different percentage of prefabs in written texts of different genres and in spoken texts produced by different speakers.

The comparison of test results of typologically different systems will reveal the universality of various mechanisms in the production of language.

3.7 Sources of Data

We can test the hypothesis on a language maximally different from English in terms of word order flexibility. The Russian language seems to be a good candidate for this study. Russian has a rich noun and verb morphology (e.g., case and tense systems) and the flexibility of the word order that can be viewed as unrestricted. Thus, Russian falls into a category of languages whose native speakers acknowledge that all logically possible constituent orders are grammatical.

The material I analyzed in the pilot study consists of a radio interview of a prominent political analyst conducted by a well-known journalist, which contains 4, 000 words[1] and 10 extracts of 100 to 400 words from different text types available on the web representing written Russian, plus two 400-word extracts from two versions of "Goldilocks". Types of prefabs that have been analyzed are lexical, grammatical and pragmatic.

The fourth type of reducibles that is used in Erman and Warren's study has been eliminated because this type of abbreviation is not practiced in Russian.

---

[1]Special thanks to Vsevolod Kapatsinski for providing the spoken data, which can be found on www.unm.edu/~alator/corpus/htm.

The number and types of Russian texts used for the pilot study were matched to the English texts in Erman and Warren (2000). The purpose of selecting this number of words and these types of written texts for the pilot study is to ensure the comparability of the results from between the two languages.

For the expanded study described in the next section, I increased the amount of data in the following way:

(1) The interviews with ten more native Russian speakers (1,000 words each) were analyzed to examine the variability among the speakers.

(2) Ten more written texts representing various genres (1,000 words each) were analyzed to check the variability among genres in written text.

The expanded data helped further to test the hypotheses put forward in this study. The issue of variability in genres and speakers surfaced in the pilot study but was not controlled for specifically. The expended study controlled for genre and speaker variation.

3.8 Results of the Pilot Study

The first question the study addressed concerned the average proportion of prefabs in spoken and written discourse. In this study, a slot represents a word. And we are interested in how many slots are filled with portions of prefab. As we can see from Table 1, the overall number of prefabs in spoken and written Russian texts is 51%, which is comparable to the overall number of prefabs in English texts of 55%.

Table 1: Proportion of prefabs in analyzed Russian texts

	Slots	Slots filled by prefabs	% of prefab slots
Spoken	4,000	1,842	46.1%
Written	4,800	2,649	55.2%
	8,800	4,491	51%

Table 1 also reveals that the density of prefabs is greater in written rather than in spoken language (55.2% vs. 46.1%). This is different from the English data, where the spoken language has a higher percentage of prefabs than the written text (59 vs. 52 %). The difference can be attributed to the variability in the use of prefabs among writers and speakers, the types of texts or language type. These results are preliminary and require further testing with a larger sample of texts that represent a variety of genres of written and spoken language. Figure 1 illustrates the comparison of the total number of prefabs in English and Russian in both written and spoken media. The English data used for comparison is from Erman and Warren (2000).

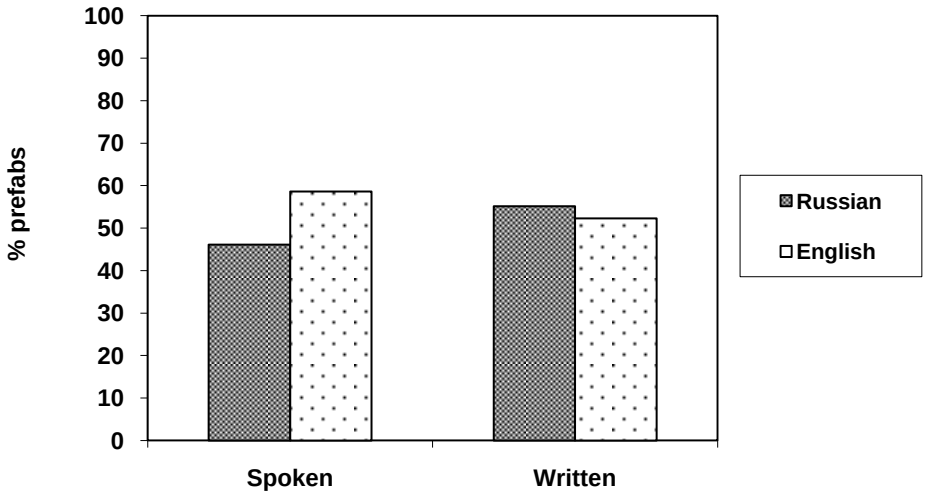

Figure 1: Proportion of prefabs in discourse

One thing that is noticeable in the English data is that Erman and Warren do not comment on the fact that the number of prefabs used in a particular text is sensitive to the genre of the written text. As I mentioned earlier, in the group of written texts I had two abstracts of the fairy tales, which had a very high density of prefabs, about 15 % higher than other written texts. I suspect that the same might be the case with the spoken language. I believe some variability exists among speakers in a community, because a chunk may or may not gain the status of a prefab, depending on the speaker.

Spoken data in the pilot study does not reflect that because all of the 4,000 tokens are taken from the same interview with the same two participants. (Some prefabs are known to all members of a linguistic community, while others are less prominent. Thus, conventionalization of prefabs is a gradual process, where co-occurrence and fusion of elements is a matter of a degree.) The difference between the written and the spoken medium is still not as great as expected and sometimes assumed (See Figure 1 for the comparison of the English and Russian data).

If we consider the distribution of prefab types, we find more striking differences: Table 2 shows we have almost twice as many lexical prefabs in written text as in spoken, which is similar to the English data (38.8% vs. 71.5%) ; considerably higher number of pragmatic prefabs in speech (25.5% vs. 5.5%) than in writing, which is what's expected. The only significantly different result from the English data is the number of grammatical prefabs in both spoken and written language (39.2% vs. 26.4%). (In English spoken, 16.7% vs. written 2.4%.) I believe this particular factor might have to do with the coding of grammatical case in Russian (the phenomenon not present in English), and other notions such as mood, aspect and tense, which are coded morphologically and constitute a greater number of grammatical prefabs in Russian than in English. Grammatical prefabs are intra-linguistic text-forming items. The high percentage of grammatical prefabs in Russian has to do with the characteristics of this morphological type of language and with what allows a greater flexibility of the word order in Russian.

Table 2: Distribution of Russian prefab types

	Lexical	Grammatical	Pragmatic
Spoken	324.8 (35.3%)	360.6 (39.2%)	234.6 (25.5%)
Written	1,008 (68.1%)	390.7 (26.4%)	81.4 (5.5%)

Figure 2, "Spoken Data", and Figure 3, "Written Data," illustrate the comparison of the distribution of lexical, grammatical, and pragmatic prefabs in English and Russian. The source of the English data is Erman and Warren (2000). The pilot study provides the data for Russian.

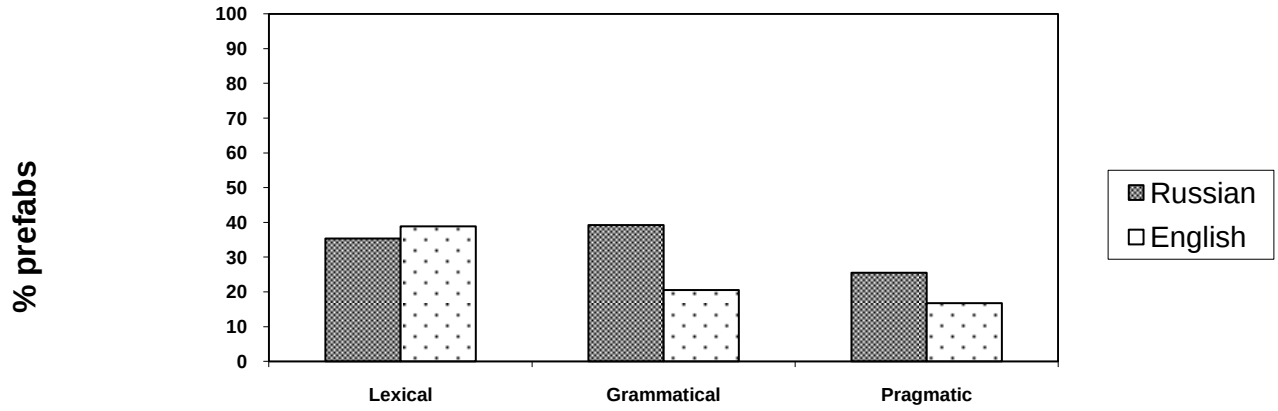

Figure 2: Comparison of spoken data

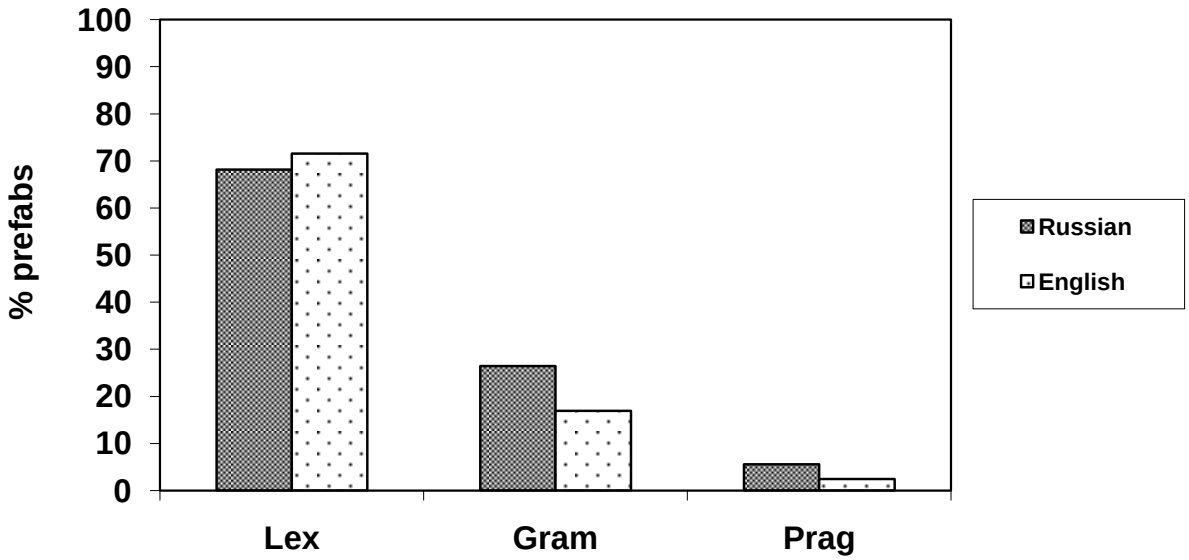

Figure 3: Comparison of written data

In addition to the distribution of various types of prefabs, the average length of prefabs was considered. The pilot study revealed that prefabs are on average shorter in spoken language than in written; the findings are presented in Table 3. The comparison of

the English and Russian data is presented in Figure 4 below. This is consistent with the English data.

Table 3: Average length of Russian prefabs

Spoken	2.58 word/prefab
Written	2.87 word/prefab

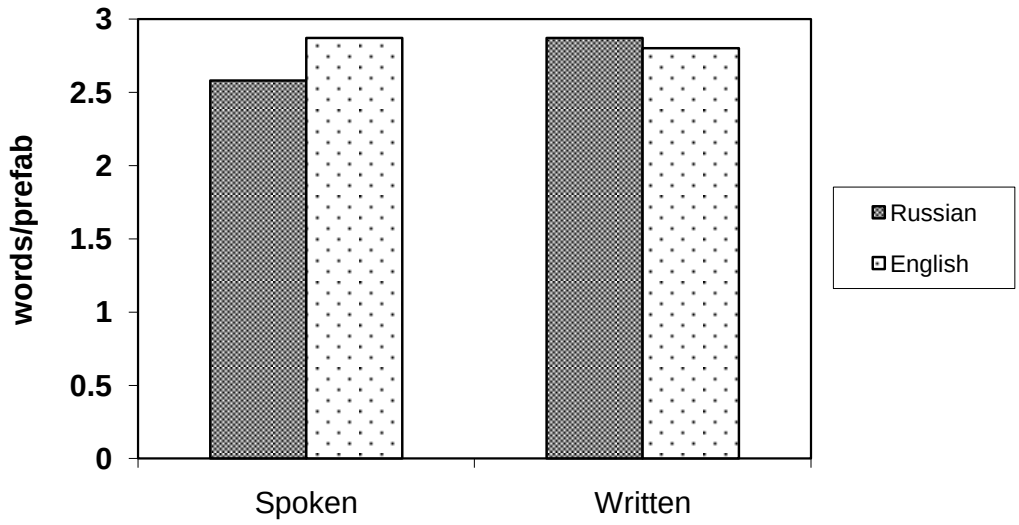

Figure 4: Comparison of average length of English and Russian prefabs

3.9 Conclusions of the Pilot Study

Results of the pilot study revealed the following:

1. A language with free word order contains on average 45 - 55% of prefabs in written and spoken discourse.
2. Higher number of grammatical prefabs in a free word order language is due to coding of grammatical case and other notions present in this morpho-syntactic type of languages.

3. Speakers of Russian use the idiom principal as much as English speakers do, thus alternating between the use of the idiom principle as well as the open-choice principle in the language production.
4. Russian data strengthens the hypothesis that the idiom and the open-choice principles are not language-specific but general principles of language production grounded in human cognition and conceptualization, which allow the minimization of processing costs and increase in fluency of the speakers.
5. One cannot make any definite conclusions about language universals based on the study of two languages. So more cross-linguistic studies of typologically different languages are necessary to advance this hypothesis and strengthen this line of argumentation.

3.10 The Current Study

3.10.1 Overview of the issues

The two principles - idiom and open choice – have been claimed by both Sinclair (1991) and Erman and Warren (2000) to be universal principles of language organization based on the common architecture of human cognition and common ways of social interaction.

Before we come to any conclusions on the universality of the open-choice and idiom principles as strategies of language processing, production, storage and use, I would like to suggest that the cross-linguistic testing of these proposals is necessary in order to investigate the effects of particular *types* of languages on the formation and usage of prefabs. Since we do not have one universal language, but rather many individual languages with many differences, the question in the forefront of typological

research, psycholinguistics, language acquisition and other research areas is just how important these differences are for the linguistic representation, processing, acquisition, and use.

The study by Erman and Warren (2000) is revealing but is nevertheless limited by the choice of a single language as the data source. English is a language that has a fixed word order that does not allow wide flexibility of ordering of elements within a clause. One might reasonably argue that fixedness of the elements in a clause contributes to the co-occurrence of elements and the formation of prefabricated strings.

It is reasonable to argue, then, that the rigidity of the English syntax might account for the high percentage of prefabricated sequences in texts and the preference of the English speakers for the idiom principle in the construction of utterances. Under such a proposition, the idiom principle becomes a language specific characteristic pertaining to English rather than a universal principle of language production. We may anticipate the presence of multiword, preconstructed units in all languages as a consequence of cognitive processing which are operative in language use -- and not its causes (Wray 2002). However, the question of how the alternation of the idiom principle and the open-choice principle manifests in various types of languages remains an open empirical question. Based on this observation, it is reasonable to hypothesize that the percentage of prefabricated strings in language with a free word order might be different. Flexibility of the word order in a language provides its speakers with a greater number of choices in coding and decoding of utterances. Thus, the speakers of a free word order language may have a greater tendency to use the open-choice principle rather than the idiom principle in language production. This issue needs to be examined before any conclusive remarks are

made on the universality of one principle or the other the nature of the alternation between the two principles, as well as, the manifestation in various language types.

3.11 Data and Coding

The current research investigates the proportion of prefabs in spoke and written Russian texts. Speech and writing are different modalities that are difficult to compare directly, because discourse emerges in specific situations where language constitutes *genres*. Genres are understood here as text or discourse types such as informal conversations, narratives, or academic essays, cultural units that are constituted by lexical and grammatical resources (Halliday and Hasan, 1989, Martin, 1992.) Different configurations of these lexical and grammatical resources can be characterized as registers. Different register choices are more or less appropriate, or more or less effective, in the realization of particular genres. The Erman and Warren (2000) study, as well as the pilot study, take into account some variability across written texts but do not control specifically for various genres. A variety of texts have been analyzed in these two studies. Specifically, fairy tales were a part of the written corpus, to represent a genre, very formulaic in form and function. In the pilot study, fairy tales proved to be much higher prefab density texts than all other texts.

Recent analyses of speech and writing have illuminated the differences in register that are reflected in different genres (Chafe, 1985; Halliday, 1987, 1989; Martin, 1989). Language users draw on different grammatical and lexical resources in creating texts of different types, and study of the different choices that are appropriate for particular genres are very informative. Prefab analysis can be an additional tool in the study of the

differences between spoken and written discourse and give us a better understanding of the role of register differences in the text production.

For the current analysis, the written corpus that was used is the Uppsala Russian Corpus (Lonngren, 1993; Maier, 1994.) The Uppsala Russian Corpus consists of some 600 Russian texts with a total of one million running words (word tokens), equally divided between informative and literary prose. The informative texts are from between 1985 and 1989, while the literary texts cover a longer period, 1960 to 1988. The corpus does not include poetry or drama. Four written genres were coded as distinct text types in the current study: 1. Fiction; 2. science-popular writing; 3. writing of social and political journalism; 4. fairy tales. Fiction and fairy tales represent literary texts, while science-popular writing and journalism writing represent informative texts. Each of four groups contains five texts of 1,000 tokens averaging to 5,000 tokens per group. The entire written corpus analyzed in this study contains 20, 000 running words (word tokens). Each text has been analyzed using Erman and Warren's method where the number, type and proportion of prefabs were calculated. In addition, a number of lexical and overall choices made by the speakers have been calculated. The results are presented in Chapter 4, Section 2, and are discussed in more details in Chapter 5, Section 2.

The spoken corpus from the pilot study was expanded by 40,000 tokens for the current study. In addition to 4, 000 tokens of spoken data obtained from the interviews available online, 40,000 tokens of spoken data came from recordings of the naturally occurring informal conversations, which consist of two to four participants each. The spoken data came from the region of Russia called Nizhnii Novgorod and included 36 participants, 13 males and 23 females. The age of participants ranged from 14 to 73 and

included various levels of education and professional development. The recorded data was transcribed using Erman and Warren (2000) conventions. The total of 44,000 running words (word tokens) were used for the analysis.

Prefab analysis was performed using Erman and Warren's (2000) method, and the number of lexical and overall choices that were made by the speakers was calculated. The results are presented in Chapter 4 and discussed in Chapter 5.

3.12 Summary of Chapter 3

Chapter 3 presents the description of the methodology developed by Erman and Warren (2000) for the study of the English prefabs. The method quantifies the amount of formulaic language used and presents a vivid picture of the interaction of formulaic and non-formulaic language in spoken and written discourse. In addition, this methodology allows us to calculate the average length of prefabs vs. non-prefabs in running texts. The length of prefabs is associated with the measure of choices made by the speaker in producing a text. Thus, there is a relationship between the number and type of prefabs (fully lexically specified or partially schematic) and the number of choices made in producing a text, i.e. a measure of estimated cognitive load involved in the process.

This methodology is used for the study of the Russian prefabs in the current research with some modifications. Russian is used for the current analysis specifically to provide a necessary contrast to English in word order type and test the put forward hypotheses of the study. Types of different prefabs in English and Russian are discussed in details. The pilot study presents the application of the method to the Russian data, results of the prefab analysis based on the small spoken and written Russian corpora, and the preliminary conclusions based on the pilot study. In the full study I expand both

spoken and written corpora and propose to specifically code for various genres that represent distinct registers and text types. The spoken data is also expanded by 40, 000 word tokens, by the number of participants and by a number of genres (in addition to the original interviews, naturally occurring informal conversations have been used).

The results of the current study for both spoken and written corpora are presented in Chapter 4. A more detailed discussion of the results is presented in Chapter 5.

CHAPTER 4

Results

4.0 Introduction

Chapter 4 reports on the analysis of prefabs and the analysis of choices in the spoken and written Russian discourse. The following research questions were addressed in this study: 1) the average proportion of prefabs in spoken and written Russian discourse (Sections 4.1- 4.1.4); 2) the distribution of prefab types across various spoken and written genres (Sections 4.2- 4.2.3); 3) the length of prefabs and non-prefabs in spoken and written corpora (Sections 4.3- 4.3.3). In addition, the analysis of choices in processing language was conducted; that is, the analysis of the number of choices made in the production of spoken and written Russian discourse (Sections 4.4- 4.4.2.1). Lexical choices, representing greater cognitive effort as opposed to grammatical or pragmatic choices, were considered separately. In both the prefab analysis and the analysis of choices, the text type was controlled for. The spoken and written media were considered separately. For each question raised, the total of six genres were analyzed. In the spoken medium, two genres were studied: spontaneous conversations and interviews. In the written medium, four genres were analyzed: fiction (FC); popular-science writing (PSW); writing of social-political journalism (SPJ); and fairy tales (FT). Statistical significance was measures using goodness of fit and contingency table testing. The chapter concludes with the summary of the results (Section 4.5).

4.1 Prefab Analysis: Proportion, Distribution, and Length

The prefab analysis of a text includes several consecutive steps. The first step in the analysis is to determine the proportion of prefabs in corpora, i.e., the number of slots filled with portion of prefabs in a text. In this method, every word in a text is considered to be a slot. By tracking the type of material used to fill the slots, we can identify how prefabs interact with each other and with non-prefabricated strings. In the prefab analysis, open slots and slots with restricted variability are not considered part of prefabs and are ignored.

The second step is to determine the distribution of prefab types across texts. The analysis of the distribution of lexical, grammatical, and pragmatic prefabs across texts allows us to document the distinction not only between speech and writing but also to document more detailed distinctions among various genres of speech and writing. Three-way classification of prefab types was adopted. The choice of this particular classification is discussed in Chapter 3, Section1.1. Potentially, this classification can be expended with more detailed groupings. However, for our purpose, to document the presence of formulaic language in Russian, this classification is adequate.

The third step is to determine the average length of prefabricated strings in a text, i.e., the number of words per prefab. This measure, in conjunction with the analysis of choices made in producing texts, gives an estimate of the effort involved in producing texts. The length of non-prefabricated strings is also analyzed. Since prefabs and non-prefabs represent different types of processing, their length has correlation with the number of choices made and thus can reflect the cognitive load of a certain type of processing. Prefabrication is seen here as a graded phenomenon, which reflects the

process of conventionalization and the gradual effects of language being used by its speakers. Again, the special interest lies in variation between spoken and written media as well as across various spoken and written genres. Let us now turn to the analysis of prefabs in the Russian corpora, where it is hypothesized that the free word order might have an impact on the structure and use of prefabs.

4.1.1 Statistical analyses

The statistical analyses of the data include 'goodness-of-fit' testing. The chi-square test for the goodness of fit of the observed distribution was performed. In order to assess whether the results were significant or not, I applied the chi-square test. This test compares the observed data with the expected data in order to assess whether the deviations are the result of chance or due to other factors. In the following discussion, I provide the results of the chi-square tests as follows: If the probability, or *p* value, for the calculated chi-square is $p > 0.001$, then the null hypothesis is maintained. The null hypothesis is generally accepted to mean there is no difference in the distributions. That is, the deviation is small enough that chance alone accounts for it. If, on the other hand, the *p* value for the calculated chi-square is $p < 0.001$, then I conclude that the results are significant, that is, a factor other than chance must be operating for the deviation to be so great.

4.1.2 Proportion of prefabs in the Russian corpora

The first question this study addresses concerns the average proportion of prefabs in texts: If every word in a text represents a slot, how many slots are filled by portions of prefabs. As was described in Chapter 3, slots filled by portions of non-prefabs are replaced with dashes while the prefab slots remain to be filled with words. This gives a

vivid illustration of the alternation between prefab and non-prefab strings. Open slots in prefabs, which can be filled by an unlimited number of words, are ignored when the number of slots filled by parts of prefabs is counted. Even if the variability is judged to be restricted, as in /go to *seminars*/ -- *seminars, lectures, classes, meetings*, etc. – it is ignored in the prefab analysis as well (but not in the estimating number of choices; see Section 4.4 below). Results are summarized in Table 1. The proportion of slots filled with parts of prefabs in spoken Russian is 64.6%, and in the written language, 58.3%. The average, according to this analysis, is 62.7%. Table1 illustrates that the percentage of prefabs in spoken language is greater than in written language. Overall, the difference between spoken and written modes is about 6%. This is less than expected or usually assumed.

Table 1: Proportion of prefabs in the analyzed texts

	Slots	Filled with prefabs
Spoken	48,469	31,311 (64.6%)
Written	20,540	11,973 (58.3%)
	69,009	43,284 (62.73%)

Functions of prefabs vary in speech and writing. Without a considerable inventory of ready-made, multiword sequences, fluent speech would hardly be possible (Pawley and Syder, 1988; Langacker, 2008). While the human brain is capable of extensive memory storage, its online processing capacity appears to be limited. To circumvent the online processing demands of time and attention, speakers rely on "chunking" to allow an increase in the quantity of material processed. Time pressure and attention limits are less

restricting in writing than in speaking; offline processing allows writers to produce a greater number of novel combinations and rely less on pre-constructed multiword sequences. Therefore, the number of prefabs in the written word is less than in the spoken word. Put differently, the novel, newly combined word-for-word strings are more frequent in writing. This finding is expected in light of the communicative and cognitive pressures associated with the differences between speech and writing. The anticipated impact of the free word order language type on the proportion of prefabs is not observed. In both speech and writing, more than half of a text is filled with prefabricated material.

4.1.3 Proportion of prefabs in spoken Russian corpus

If we consider spoken and written media separately, we observe more distinctions in the use of prefabs. Two genres in spoken corpus were considered – interviews and natural conversations. Overall, in spoken language 65% of all slots are filled with a portion of prefabs. Table 2 demonstrates that the density of prefabs is greater in interviews, 66.8%, than in spontaneous conversation, 64.2%. The variation can be attributed to the difference in speaking tasks and to the difference in the verbal ability of the speakers. Interviews tend to be more scripted than spontaneous conversations. The questions for an interview are designed in advanced, and the flow of the information is guided by an interviewer. On the other hand, spontaneous conversations can be less predictable and restricting. While a conversation may be influenced by a choice of topic, associated context, background knowledge, relationships between participants and other factors, it still represents a less formal verbal exchange than an interview. The difference between the two genres, however, is not statistically significant as evidenced through ‘goodness-of-fit’ testing.

Table 2: Proportion of prefabs in spoken Russian

Genres	Slots	Filled by prefabs
Interviews	8,098	5,409 (66.8%)
Conversations	40,371	25,901 (64.2%)
	48,469	31, 311 (64.6%)

Overall, in speaking on average 65% of slots is filled with prefabs. Speakers seem to vary in their ability to use prefabs. The interviews between a journalist and a political analyst display greater fluency and a faster rate of speaking than what is observed in spontaneous conversation data.

The proportion of prefabs in texts gives a more realistic view of the lexicon. Traditional lexicon is based on a dictionary metaphor, which lists individual words (i.e. lexemes) and is generally considered a repository for idiosyncratic material. Traditionally, this mental dictionary is a list of words out of which phrases and sentences are built according to the rules of syntax. If prefabs amount to 65% of language that is stored and retrieved as whole, prefabs must be listed in the lexicon as well. Prefabs are conventional structures that are retrieved as wholes. Thus, Jekendoff (1995) states: "The number of phrases in the lexicon is even greater than the number of individual words." In light of these facts, the memory capacity for language must be extensive. Bolinger (1976: 2) repeatedly argued that we "store a large number of complex items which we manipulate with comparatively simple operations." The questions that arise are: (a) What is the unit of grammar? and (b) What are the combinatorial rules that act upon such units? In view of the fact that 65% of language is formulaic, i.e. stored and

retrieved as a whole, I suggest that the unit of grammar is a multiword sequence of any length that is stored and retrieved holistically by a speaker. The combinatorial system acts on such units and creates sequences of progressively greater complexity.

4.1.4 Proportion of prefabs in written Russian corpus

When the written texts were selected, care was taken to match the Russian text types to the ones used in the English prefab study (Erman and Warren, 2000) to ensure the comparability of the results. However, Erman and Warren do not always specify the genres of the texts they analyzed. The exception is the genre of fairy tales that is mentioned and described explicitly. The same number of texts representing the genre of fairy tales was used in the Russian written corpus. Fairy tales is a highly formulaic genre and were chosen to provide a contrast with less formulaic texts.

Four genres of written language were analyzed. In addition to fairy tales (FT), fiction (FC), popular-science writing (PSW), and writing of social and political journalism (SPJ) were added to the corpus. All four genres are well established in the Russian literary tradition and are well recognized by readers. Each genre contains five texts of about 1,000 tokens, averaging to more than 5,000 tokens per group. The entire written corpus analyzed in this study contains 20,540 tokens. Table 3 reveals the proportion of slots filled by parts of prefabs in four types of written texts.

Table 3: Proportion of prefabs in written Russian texts

Genres	Slots	Slots filled by prefabs	% of prefab slots
FC	5,214	3,071	58.9%
PSW	5,106	2,818	55.2%
SPJ	5,202	2,830	54.4%
FT	5, 018	3,254	64.9%
	20,540	11,973	58.3%

The range in the number of slots filled with portions of prefabs across written genres analyzed is from 54.4% to 64.9%. The average proportion of slots filled with parts of prefabs in written language is 58%. The highest density of prefabs is in fairy tales (FT), at almost 65%. In this study, one fifth of the fairy tales corpus is composed of prefabs wholly repeated from one place to another, which is a typical characteristic of the genre. The second genre with the highest percentage of prefabs is Fiction (FC), which contains 58.9%. Popular-science writing and socio-political journalism follow with 55.2% and 54.4%, respectively. The difference between FT and SPJ is significant ($p < 0.001$) as evidenced through chi-square testing.

4.2 Distribution of Prefab Types in Discourse

The distribution of three types of prefabs - lexical, grammatical, and pragmatic - was analyzed in two genres of spoken and six genres of written language. The goal is to identify whether specific types of prefabs serve any particular functions and whether it is possible to relate certain types of prefabs to certain types of text or modality. Additionally, this analysis reveals whether any prefabs are stylistically neutral. As

discussed in previous chapter, lexical prefabs are semantic units that represent some extralinguistic entity or phenomenon, whereas functional prefabs (i.e., grammatical and pragmatic) prefabs are intra-linguistic, text-forming units (see Chapter 3, Section 3). There are different reasons for conventionalization of various types of prefabs. Driven by the grammaticalization processes, some lexical prefabs acquire pragmatic functions and begin to resemble functional prefabs in their use. Some examples include: *that's true, it's all my fault, don't bother, is that all?, that's a good idea, what is it all about?* Thus, the distinction between lexical and functional prefabs is fluid at times. However, for practical reasons, Erman and Warren include debatable cases in the lexical category. This matter of policy was followed in the current analysis for consistency and comparability of results.

The next section provides the results of the distribution of prefab types. The differentiation between speech and writing, as well as between speakers and genres is adhered to. The distribution of prefab types across media and genres discussed more thoroughly in Chapter 5. The findings have implications for models of language structure and use.

4.2.1 Distribution of prefab types in Russian corpora

If we consider the distribution of prefab types, we find more striking differences between speech and writing. Table 4 demonstrates that there are almost twice as many lexical prefabs in the written medium (73.1%) than in spoken (37.9%). The proportion of grammatical prefabs in speech (23.9%) and writing (17.6%) is comparable with a difference about 5%. As expected, a greater number of pragmatic prefabs in speech (38.2 %) than in writing (9.3%) was observed.

Table 4: Distribution of prefab types in Russian

	Lexical	Grammatical	Pragmatic
Spoken	11,897 (37.9%)	7,427 (23.9%)	11,987 (38.2%)
Written	8,755 (73.1%)	2,102 (17.6%)	1,116 (9.3%)

The difference between spoken and written medium in distribution of prefabs is significant. There are nearly twice as many lexical prefabs in writing than in speaking. This is puzzling because some difference between the two modes is expected but not quite to this extent. More lexical prefabs are expected in writing than in speaking due to the features associated with the written medium, specifically, the lack of time pressure and a lack of attention deficit, which are strong factors present in on-line communication. More time allows writers to consider more choices and make a variety of selections. Nevertheless, the difference of 35.2 % is striking and puzzling.

Next, the range of grammatical prefabs between speech and writing is 6.3%. Grammatical prefabs are text-forming by definition rather than referring units. This means that grammatical prefabs serve the functions of modifying, quantifying, and specifying the reference or meaning of nouns, verbs, adjectives, or adverbs. It has been pointed out that they often serve as their substitutes or as links between propositional or referring items. According to our analysis, there are more grammatical prefabs in speech than in writing.

The last group of prefabs I considered was pragmatic prefabs. The number of pragmatic prefabs is significantly higher in speech than in writing, specifically, by 28.9%.

This is expected due to the nature of pragmatic prefabs, which serve text-monitoring, interpersonal, or metalinguistic functions. Pragmatic prefabs are functional in that they do not directly partake in the propositional content of the utterance in question. Most pragmatic prefabs are restricted to spoken language, and some have functions that could be expressed in writing by punctuation, paragraphing, or in other orthographic conventions. Pragmatic prefabs are often multifunctional. They serve not only different functions in different contexts but can serve different functions in the same context. Pragmatic prefabs are unique in this respect, because grammatical or lexical prefabs tend not to be multifunctional. Also, pragmatic prefabs occur almost exclusively in speech or reported speech. However, pragmatic prefabs tend to be more like grammatical prefabs rather than lexical ones; they tend to be short and relatively invariable and appear to be restricted in number. It is reported for the English data that the type-token ratios of the three groups are 1000-980 for lexical, 1000-650 for grammatical and 1000-660 for pragmatic prefabs (ibid: 45). The Russian type-token ratios for the three groups seem to follow the pattern - 1000-900 for lexical, 1000-700 for grammatical, and 1000-750 for pragmatic prefabs.

4.2.2 Distribution of prefab types in spoken Russian corpus

The distribution of the three categories of prefabs – lexical, grammatical, and pragmatic – was analyzed in two spoken genres of Russian, which are the interviews and the spontaneous conversations.

Table 5: Distribution of prefab types in spoken Russian corpus

Genres	Lexical	Grammatical	Pragmatic
Interviews	2,288 (42.3%)	1,363 (25.2%)	1,757 (32.5%)
Conversations	9,609 (37.1%)	6,060 (23.4%)	10,230 (39.5%)
	11,897 (37.9%)	7,427 (23.9%)	11,987 (38.2%)

The verbal skills vary across speakers in various ways. Overall, the average number of lexical prefabs is 37.9%; grammatical, 23.9%; and pragmatic, 38.2%. The number of lexical prefabs is higher in interviews (42.3%) than in conversations (37.1%). This difference can be due to either genre or speakers. Grammatical prefabs are almost equal between the two genres (no significant difference). The greater number of pragmatic prefabs is observed with greater number in conversations (39.5%) than in interviews (32.5%). This difference in the number of pragmatic prefabs is attributed to the difference in spoken genre. The interviews are set up with predictable format. A smaller amount of prefabs associated with turn-taking negotiations is present in interviews than in conversation. Both interview and conversational data show that lexical prefabs are used frequently, with pragmatic prefabs following and grammatical prefabs a distant third. The high frequencies of prefabs in speaking are expected due to their facilitator effect enabling fluency and speed.

4.2.3 Distribution of prefab types in written Russian corpus

The distribution of three types of prefabs - lexical, grammatical, and pragmatic - were analyzed in four written genres of Russian, which include fiction (FC), popular-science writing (PSW), writing of social and political journalism (SPJ) and fairy tales (FT).

Table 6: Distribution of prefab types in written Russian corpus

Genres	Lexical	Grammatical	Pragmatic
FC	2,254 (73.4%)	556 (18.1%)	261 (8.5%)
PSW	2,004 (71.1%)	486 (17.3%)	328 (11.6%)
SPJ	2,063 (72.9 %)	504 (17.8%)	263 (9.3%)
FT	2,434 (74.8%)	556 (17.1%)	264 (8.1%)
	8,755 (73.1%)	2,102 (17.6%)	1,116 (9.35%)

The greatest number of prefabs in written text is lexical (73.1%), followed by grammatical prefabs (17.6%), and pragmatic, (9. 35%). The range for lexical prefabs across various genres of written text is 3.75%, with the highest number for fairy tales 74.8%, and then in fiction 73.4%. The lowest frequency is in PSW 71.1%, followed by SPJ at 72.9%.

The second group of prefabs in terms of frequency of occurrence in writing is grammatical. The range for grammatical prefabs across written genres is 1%, which is not significant. The highest number of grammatical prefabs is in fiction at 18.1 % and the lowest is in Fairy tales at 17.1%. In the middle of the range is PSW at 17.3% and SPJ at 17.8%.

Pragmatic prefabs are not frequent in writing in comparison with speech; the range across written genres is 3.5%. The highest number of pragmatic prefabs is in SPW at 11.6%, and the lowest in fairy tales at 8.1%. Fiction is at 8.5% and SPJ at 9.3%. Lexical prefabs are most prominent in writing, with grammatical prefabs to follow and pragmatic being the third group in terms of frequency.

If we compare the distribution of prefab types between spoken and written media, the differences are striking. The number of lexical prefabs in writing is almost twice as high as in speaking (73.1% vs. 37.9%). The average proportion of grammatical prefabs in speech is 23.9% and in writing, 17.6%, for a difference of 6.3 percent. The number of pragmatic prefabs is significantly higher in speech, 38.2%, than in writing, 9.4%, with the difference at 28.9%. The data clearly show that the distribution of prefab types varies significantly greater along the speech-writing continuum, rather than along speaker and genre variation.

4.3 Length

Length of prefabs is a measure of the number of words per prefab. This measure reflects the mechanism of chunking involved in language production. In conjunction with the analysis of choices, this measure gives an estimate of the processing effort involved in production of texts. The main questions are: (a) How do prefabs combine with each other and with non-prefabricated strings? (b) Is this alternation affected by the free word order type of language?

4.3.1 The length of Russian prefabs

The length of prefabs in both spoken and written corpora has been considered. We find a difference in that prefabs are on average somewhat shorter in spoken language that in written, which is evident in Table 7. The average length of prefabs in speech is 3.25 words/prefab, while in written texts it is 4.12 words/prefab. The difference averages to 1.13 words/prefab between speech and writing. These results warrant a closer inspection of prefab types and their length.

Table 7: Average length of prefabs

Spoken	3.25 words/prefab
Written	4.12 words/prefab

If we consider the average length of prefab types, we can see from Table 8 that the lexical prefabs are longer than other types of prefabs. According to this model of analysis, the length of lexical prefabs varies between three and six words per prefab. Some prefabs are longer than six words, but they are relatively few. The average length of lexical prefabs is 4.03 words per prefab. They are followed by grammatical prefabs with 3.89 words per prefab. The shortest prefabs are pragmatic at 2.98 words per prefab. Considering the fact that pragmatic prefabs occur almost exclusively in speech, and prefabs in speech tend to be shorter than in writing, the fact that pragmatic prefabs are the shortest is not surprising.

Table 8: Average length of prefab types

Lexical	4.03 word/prefab
Grammatical	3.89 word/prefab
Pragmatic	2.98 word/prefab

A more detailed analysis of prefab types and their length was considered for the interview and conversational data (Section 4.3.1) as well as for the four genres of the written texts (Section 4.3.2).

4.3.2 The length of prefabs in spoken Russian corpus

As noted in Section 4.1, the number of slots filled with portion of prefabs is higher in spoken rather than written discourse. However, if we examined the length of prefabs in two media, we noticed that the length of prefabs in spoken texts is shorter. Table 9 reveals the length of various types of prefabs in interviews and in natural conversation. Lexical prefabs average 3.51 words/prefab in the interview data and 3.55 words in conversational data. Grammatical prefabs tend to be shorter than lexical or pragmatic prefabs in both genres - 2.89 words/prefab in interviews and 2.91 in conversation. Pragmatic prefabs that are extremely high in both interviews and conversations average at 3.59 and 3.31 words/prefab, respectively.

Table 9: Average length of prefabs in spoken corpus (number of words per prefab)

	Lexical	Grammatical	Pragmatic
Interviews	3.51	2.89	3.59
Conversation	3.55	2.91	3.31

Many researchers advocated that intonation unit (IU) – a basically prosodic entity – is a relevant unit of spoken language. IUs are found to be longer in written than in spoken language, a fact that was attributed to different processing constraints. It has been argued by Chafe and Danielewics (1987) that spoken intonation units are limited in size by the short-term memory of focal consciousness capacity of the speaker, and perhaps also by the speaker's awareness of the listener's capacity limitations. Writers of cause are under no such constraints, and as a consequence they have a tendency to produce expanded IUs. To my knowledge, the relation between IUs and prefabs has not been

studied. However, the results from both areas point in the same direction - shorter units in speaking than in writing. The explanation is attributed to the difference in processing constraints between the two modes. The different ways prefabs are joined together and how they are joined to newly constructed language to form sentences in spoken and written language is worthy of observation. Spoken language relies to a large extent on a chaining technique and avoids elaborate syntactic relations among clauses. It has been argued that IUs are the natural unit of speaking, whereas integrated, elaborated sentences have become the natural unit of writing. The use of prefabs exhibit similar tendencies.

4.3.3 The length of prefabs in Written Russian Corpus

Prefabs tend to be longer in writing than in speech; the non-prefabricated units are also longer in writing than in speech. Results are summarized in Table 10. The average length of prefabs in writing is 4.12 words/prefab. The range is from 4.21 lexical prefabs in FC to 2.7 grammatical prefabs in SPW. The longest prefabs are lexical; the average length of lexical prefabs is 4 words/prefab. In FC we see 4.21 words/prefab; in FT 4.19; PSW at 3.81 and SPJ at 3.78 word/prefab.

Pragmatic prefabs are the second group in length measure. The average length of pragmatic prefabs is 3.9 words/prefab. Variation includes the longest prefabs 4.3 words/prefab in fairy tales and very close measures in the other 3 genres - 3.9 word/prefab in fiction, 3.7 in PSW, and 3.8 in SPJ.

Grammatical prefabs form the third group. The average length is 2.83 words/prefab with 2.8 words/prefab in fiction, 2.7 in SPW, and 2.9 in both SPJ and fairy tales. Both Tables 10 and 11 reveal that the length of prefabs contrasts between speech and writing to a greater extent than across genres or speakers.

Table 10: The average length of prefabs in written Russian corpus (words/prefab)

	Lexical	Grammatical	Pragmatic
Fiction	4.21	2.8	3.9
SPW	3.81	2.7	3.7
SPJ	3.78	2.9	3.8
FT	4.19	2.9	4.3

We can look at the length of prefabs and non prefabs in another way.

Table 11: Average length of non-prefab and prefab strings

	Non-prefab	Prefab
Spoken	2.89/2.43	3.25/3.19
Written	4.31/3.01	4.12/4.30

Two figures are reported under non-prefab in Table 11. The first figure refers to non-prefab strings indicated by dashes in coding; the second to non-prefab strings including adjacent open slots and open slots occurring within prefabs, which are short. The first figure under prefab refers to single prefabs; the second to strings of successive prefabs. The findings reveal that both prefab and non-prefab strings are longer in written than in spoken language, the effect being particularly significant in non-prefabricated strings. The study reveals that the most important difference between spoken and written languages is the length of non-prefabricated strings in writing; the non-prefabricated strings are significantly longer in writing than in speech. The difference in processing

constraints associated with each medium – speech and writing – contributes to the difference in the results.

In a running text, prefabs can combine with each as well as with the strings of non-prefabs. Thus an elaborate web is constructed. The interaction of prefabricated language with the novel language is the important question this study addresses. Table 11 reveals the length of non-prefabricated strings in texts. The typical range for non-prefabricated strings is two to six slots. The majority of non-prefabricated strings are only one or two words. This is especially evident in spoken language, where cognitive limitation of online processing is in effect. Non-prefab strings in both written and spoken language of more than 10 words are rare. A generalization is that the longer the non-prefabricated combination, the rarer it is. If we disregard poems or memorized literature, it is clearly the case that the frequency of multiword strings decreases with length. Thus, full texts might occur only once, some sentences many times, and some multiword strings occur more frequently than many words. Do the frequencies of these sorts of language structures play a role in processing? If yes, is that role comparable to the role of frequency in lexical processing?

4.4 Choice

The second portion of this method is the analysis of choices made in producing a text. A choice represents a cognitive effort made by a language user in encoding and decoding a message. The higher is the number of choices that needs to be made, the greater is the effort for processing language. Traditional assumption in the generative model is that each position in a clause offers a choice. What the current method shows, however, is that the presence of prefabs in a clause reduces the number of choices a

speaker must make in constructing a language. Retrieving each word separately from the mental storage represents a greater cognitive effort than choosing a preconstructed, ready-made sequence of multiple words. Such examples as *What's up?, How are you doing?, Good to see you,* which are routinely used as greeting formulas, represent a single multiword retrieval from the mental lexicon of a speaker. Typically a prefab of multiple words represents a single choice. However, because prefabs can have some flexibility with open- and restricted variability slots, they can represent more than one choice. As was noted earlier, the length of prefabs may vary according to medium, speaker, genre, and possibly other factors with the average length of prefabs of 2.8 and 3.5 in the English and Russian corpora, respectively. The longest prefabs documented in both corpora were 12 and 14 words per prefab, even though more typical examples were two to four words. The presence of prefabs significantly reduces the number of choices made, and thus it reduces the cognitive effort on the part of the speaker.

It has been mentioned that prefabs with open slots or slots with restricted variability can represent more than one choice. Both an open-choice slot and a restricted-variability slot represent additional choices to be made. When the number of choices made in constructing a text is calculated, open and restricted variability slots are counted. However, due to the presence of prefabs in any text, whether written or spoken, produced by any language user of any genre, the number of retrievals is always less than the number of words in a text due to the presence of prefabs. If more than half of a text consists of prefabricated language that the number of choices made in constructing a text is reduced. The English data suggests that in a text of 100 words, with 55% of slots filled by portions of prefabs, on average only 45 single-word choices would be made. This

figure does not tell us how many choices there are in total in a text. A spate analysis of choices gave us the average figure of 71 choices in 100 words, which is a combined figure of the number of single-word and prefab choices. If we compare the figure for single-word choices (45) with the figure for the total number of choices (71), we realize there are only 26 choices instead of the 55 there would have been had there been no prefabs. Of this 26, some constitute restricted choices within prefabs requiring comparatively little mental effort.

This statistic demonstrates that the number of choices in both spoken and written corpora is significantly lower than a number of slots per corpora. The presence of prefabs in any text reduces the cognitive effort involved in processing and therefore increases speed and fluency of the speaker. The number of choices made gives some idea of the processing effort involved in composing a text.

Consequently, the difference between counting words/slots and counting choices in any text is significant. Prefabs being chunks of automated speech represent one single choice, whereas non-prefabs represent a novel language, with each slot being a separate choice. The use of prefabs in any given text means reducing the number of choices a language user has to make because a prefab by definition is a combination of at least two orthographic words. As demonstrated in Section 4.3, the length of prefabs varies due to the medium, speaker, and genre of the text with an average of 3.5 words per prefab in speech and 4.12 in writing. Studies show that the number and length of prefabs can vary due to the medium used, topic discussed, and speakers involved. Multiple variables can influence prefab proportion, distribution and length. In addition to the ones considered in

this study, social factors of the speakers such as education, background, age, gender, profession and others may influence the use of prefabs.

4.4.1 Number of choices in Russian discourse

All modalities, text types, and genres contain prefabs, which constitute single multiword retrievals from our mental lexicon. Because of prefabs, the number of retrievals is fewer than the number of words in a text of any size. However, determining the number of prefabs in a text does not automatically reveal the number of retrievals due to some variability inherent in prefabs. The analysis of choices was designed to reveal the number of choices in a text. This analysis of choices is indicative of the processing effort involved in composing a text, and thus, it allows comparison between text types, genres, speakers and languages.

The method is designed in such a way that only choices of slot-fillers/words are considered. Choice of tense, mood, number for the verb, nominative form for the subject, or accusative form for the object are not considered as a matter of policy. In other words, in this analysis, one slot can never represent more than one choice. Thus, tense-forming auxiliaries do not represent separate choices. Table 12 summarizes the number of choices in Russian texts. Overall, 65% of all slots represent a choice in speaking and 78% of all slots represent a choice in writing.

Table 12: Number of choices in spoken and written language

Spoken	Written
65%	78%

As expected, there are more choices made in composing a written than a spoken text. This difference is statistically significant, ($p < 0.001$) as evidenced through chi-square testing. The numbers represent the overall number of choices made including single-word, open and restricted variability, and prefab choices combined. There is a correlation between the number of prefabs and the number of choices made in a text. The higher the number of prefabs the lower the number of choices and vice versa.

4.4.2 Choice in spoken Russian corpus

Analysis of choices in two genres of spoken Russian was conducted. The two spoken genres were interviews and spontaneous conversations (See Chapter 3, Section for details). Table 13 demonstrates that the overall number of choices made in interviews is 63% and in conversations 67% of all slots.

Table 13: Number of choices in spoken Russian corpus

Interviews	Conversations
63%	67%

The variation between interviews and spontaneous conversations is 4%. This is not statistically significant. Fewer choices were made in interviews than in conversations. Interviews are more pre-constructed and less spontaneous than conversations. The interviews were conducted between skilled speakers, specifically, a journalist and a political analyst. Both parties in the interviews displayed mastery in their use of prefabs as reflected in their fluent and fast pace exchange. Consequently, the number of choices they had to make was reduced. This combination of reduced number of choices and an increased number of prefabs is undoubtedly what contributes to the perceived fluency,

speed, and mastery of the spoken language by both parties. Similar to the way some people are naturally more verbal than others, some speakers, due to their professional engagements, natural ability, or extensive practice, can be better users of prefabricated multiword sequences of a language.

The transcript of conversations revealed the presence of many hesitations, pauses, conversational repairs, and turn-taking negotiations. Human capacity for encoding novel speech while speaking appears to be limited. By observing participants in a conversation, we see an attempt by the speakers to overcome this limitation by using prefabs. While prefabs are present in conversation, the conversational style among speakers seems to be more disjointed and less fluent than in the interviews. The difference in the number of choices between the interviews and conversation is attributed to the professional skills of the speakers involved in the interviews.

4.4.2.1 Choices in written Russian corpus

There are fewer prefabs and more choices in writing than in speaking. Some variation among written genres was observed. Table 14 summarizes the results of the analysis of choices in four written genres – fiction (FC), popular-science writing (PSW), socio-political journalism (SPJ) and fairy tales (FT).

Table 14: Number of choices in written Russian corpus

FC	PSW	SPJ	FT
73%	76%	78%	71%

The density of prefabs in highly formulaic genres such as fairy tales contributes to the fewer choices made in production of texts. Less formulaic genres exhibit fewer

number of prefabs and higher number of choices made. Thus, in written Russian texts, the greatest number of choices was made in writing of social and political journalism (SPJ), 78%, followed by popular-science writing (PSW) 76% of choices made. Fiction (FC) was the third group with 73% followed by fairy tales (FT) with the lowest number of choices made 71%. The range of variation among genres is statistically significant for the p value previously defined. There is a correlation between the choices made and the number of prefabs present in any particular genre. For instance, a fifth of the fairy tales corpus is composed of prefabs wholly repeated from one place to another. In highly formulaic genres such as fairy tales with frequent repetitions of prefabs, the number of choices made is reduced.

4.4.3 Lexical choices in Russian Corpus

Lexical choices represent a greater cognitive effort than grammatical or pragmatic choices. Choosing an item from a restricted set is less demanding than selecting one from an indefinitely large set. Therefore, the number of lexical choices was estimated in addition to the overall number of choices. Table 15 summarizes the results of this analysis of the spoken and written texts.

Table 15: Number of lexical choices in spoken and written language

Spoken	Written
29% of all slots	46% of all slots
41/% of all choices	58% of all choices

As expected, there are more lexical choices in written than in spoken language. The difference is statistically significant, $p < 0.001$ as evidence through chi-square testing.

As an illustration of the analysis of choices, an extract from one of the English written text samples is given below. Lexical choices are represented by a combination of underlining and a bold shrift.

(16) **To the best of** + my + **knowledge**, there is no **record of** a **society** which **has used literacy** for + the **profane** and **imaginative** + **purposes** and which **has** + not + **produced books dealing with sexual topics;**

Thus, example (16) contains 33 slots (i.e., words), 20 of which are filled by prefabs, 23 choices, 13 of which are lexical. It also has six prefabs with four lexical prefabs, one grammatical and one reducible.

It has been documented that the processing effort associated with making a choice in general and a lexical choice in particular is related to the frequency of a particular item. High-frequency words are accessed faster than low-frequency words. Various frequency effects such as the reduction effect, the conserving effect, and the effect of automatization have been documented (Bybee 2006, 2010; Hyman, 1998) and linked to the formation of prefabs.

For many years, models of lexical processing were differentiated in terms of how they accounted for the fact that higher frequency words are more readily available than low-frequency words. Another question that has not been resolved is the extent to which frequency effects are fundamentally lexical in nature. We know, for example, that idioms are likely to be stored in the mind in a manner that is comparable to the manner in which

words are stored. The main argument for this is based on their non-compositional semantics. It also has been reported that ordinary multiword sequences which are non-idiomatic and fully compositional are also subject to frequency effects (Tremblay, 2009; Boas, 2003). It is now clear that the Chomskyan claim that sentences are largely unique events does not agree with the facts (Bybee, 2006, 2007 and references cited therein.) Although native speakers can easily produce and understand sentences they have never heard, many sentences are heard repeatedly.

4.4.3.1 Lexical choices in spoken Russian corpus

Lexical choices were considered separately in interviews and conversations. Lexical choices represent an open-ended set while grammatical choices are restricted. Presumably, choosing an item from a restricted set is less demanding than selecting one from an indefinitely large set. Therefore, the method includes a separate analysis of choices classified as lexical in nature. Table 16 displays the results of the analysis of lexical choices in interviews and spontaneous conversations.

Table 16: Number of lexical choices in spoken language

Interviews	Conversations
25% of all slots 36% of all choices	27% of all slots 39% of all choices

The variation between the two genres is minimal. More lexical choices were made in conversations than in the interviews - by 2% more of all slots and 3% more of all choices. Needless to say, the difference is statistically not significant, $p > 0.001$ as evidenced through chi-square testing.

4.4.3.2 Lexical choices in written Russian corpus

Table 17: Number of lexical choices in written language

Fiction	SPW	WSPJ	FT
45% of all slots 59 % of all choices	43% of all slots 57% of all choices	40% of all slots 55% of all choices	35% of all slots 51% of all choices

4.5 Summary of Results

Chapter 4 presents the results of the prefab analysis and the choice analysis in spoken and written Russian corpus following the method of Erman and Warren (2000). The prefab analysis is includes the identification of the proportion, distribution, and length of prefabs in speech and writing (Section 4).

Proportion of prefabs addresses the question of the number of prefabs in texts as opposed to non-prefabs. Speech was found to be more formulaic than writing. Two genres of spoken language and four genres of written language were considered for each question raised. It was found that on average more prefabs are used in spoken texts than in writing. Speech tends to be highly formulaic, which contributes to speed and fluency of the speakers. Variation among spoken and written genres was found. The parties engaged in interviews used multiword prefabricated sequence slightly more often than speakers in spontaneous conversations. The difference was attributed to the professional skills of the interviewers as opposed to speakers involved in conversation (Sections 4.1 – 4.1.4).

The analysis of the distribution of prefab types revealed a more striking difference between spoken and written medium. The distribution of prefab types addressed the

question of the number of various types of prefabs - lexical, grammatical, and pragmatic across spoken and written genres. There are twice as many lexical prefabs in writing than in speaking. The number of grammatical prefabs is comparable between the two media, while the number of pragmatic prefabs is significantly higher is speech. In fact, pragmatic prefabs occur almost exclusively in speech. The difference between interviews and conversations was minimal. There were more lexical prefabs in interviews than in conversations, by 5.2%. There were more grammatical prefabs in interviews than in conversations, by 1.8%. And there were more pragmatic prefabs in conversations than in interviews, by 7%. The differences are explained in light of professional skills, goals and setting of the exchanges (Sections 4.2 - 4.2.2).

The distribution of prefab types across four written genres analyzed displays variation. The number of lexical prefabs was the highest in fairy tales (74.8%) and lowest in popular-science writing (71.1%). Grammatical prefabs were the highest in fiction (18.1%) and lowest in fairy tales (17.1%). Pragmatic prefabs are highest in writing of social-popular journalism (9.3%) and lowest in fairy tales (8.1%). The differences were within an expected range and are explained by stylistic variation among genres. The analysis allows comparison between text types and contributes to the genre studies (Section 4.2.3).

The length of prefabs and non-prefabs in texts analyzed was measured. It was found that prefabs were longer in writing than in speaking. Among spoken genres, lexical prefabs were longer in conversations (3.55) than in interviews (3.51); pragmatic prefabs were longer in interviews (3.59) than in conversations (3.31). The length of prefabs in written texts was distributed in the following way: The longest lexical prefabs

were in fiction (4.21words/prefab) and shortest in social-popular journalism (SPJ), (3.78 words/prefab); grammatical prefabs were the same length (2.9 words/prefab) in fiction and SPJ; pragmatic prefabs were the longest in fiction (4.3 words/prefab) and the shortest in popular-science writing (PSW), (3.7 words/prefab). Non-prefabs are also longer in writing than in speaking. On average non-prefabricated strings are 4.31 words/prefab in writing and 2.89 in speech (Sections 4.3 – 4.3.2).

Chapter 5

Discussion: Russian and English Prefabs Compared

5.0 Introduction

In Chapter 5, I discuss the results of the study and provide the comparison of the structure of formulaic language in Russian and English. The prefab analysis and the analysis of choices combined are based on a unique method for the study of multiword prefabricated sequences, which allows cross-linguistic comparison of text types. Originally designed by Erman and Warren (2000) for the study of English prefabs, the method is currently applied to Russian, which is a free word order language. The results of the analyses reported in Chapter 4 revealed that the use of prefabs in Russian texts is extensive, and variation in distribution of types of prefabs and their length exists along a speech-writing continuum, as well as across genres.

The current chapter discusses results of the study and provides the Russian and English comparison of prefabs. The analysis of prefabs in texts revealed that 1) there is a greater proportion of prefabs in a free than in a fixed word order language (Section 5.1.1); 2) the variation in distribution of prefab types is stronger across medium and genres than between the two languages (Section 5.1.2); 3) prefabs are longer in a free word order than in a fixed word order language (Section 5.1.3). The analysis of choices revealed that there are more choices made in a free than in a fixed word order language (Section 5.2). The variation between spoken and written medium and across genres is examined in two genres or spoken and four genres of written Russian texts (Section 5.3). The English data used for the two-language comparison comes from Erman and Warren (2000). Implications of the prefab analysis and the analysis of choices for holistic and

analytic processing are explored based on the results of the current study. The findings support a usage-based model that is emergent in the context of language use.

The role of word order and case marking in the use of formulaic language is examined in Section 5.4. The chapter concludes with the summary in Section 5.5.

5.1 Prefabs in English and Russian: Proportion, Distribution and Length

The comparison of the results of the current study with the English data revealed similarities as well as differences in proportion, distribution, and length of prefabs in the two languages. The proportion of prefabs in a text is the number of prefabricated sequences in language and correlates with the holistic mechanism of language processing, as opposed to non-prefabricated sequences, which are processed analytically. The holistic/analytic continuum accommodates all of the structures of language from completely fixed and idiomatic to the ones that are novel and compositional. Distribution of prefab types across spoken-written distinction and across genres revealed variations. Similarly, the length of prefabs varies along a spoken/written continuum and across genres. Overall, the number of choices made in English and Russian are compared.

5.1.1 Proportion of prefabs in Russian and English texts

The average proportion of prefabs in Russian texts is 62.7% compared to 55.4% in English. Both Russian and English data attest to the fact that more than half of any text in these two languages contains prefabricated, multiword sequences. The proportion of prefabs in the spoken Russian data is 64.6%, and in spoken English is 58.6%; the difference is 6%. The difference is statistically significant ($p<0.001$). The proportion of prefabs in the written data is 58.3% in Russian and 52.3% in English. Overall, results are summarized in Table 1.

Table 1: Proportion of Prefabs in Russian and English texts

	Russian	English
Spoken	64.6 %	58.6 %
Written	58.3 %	52.3 %
	62.7 %	55.4 %

These findings confirm hypothesis (1), which states that a language with a free word order will use idiom principle in the construction of sentences just as often as a language with a fixed word order, as discussed in Section 3.6. The comparison of the English and Russian data is illustrated in Figure 1:

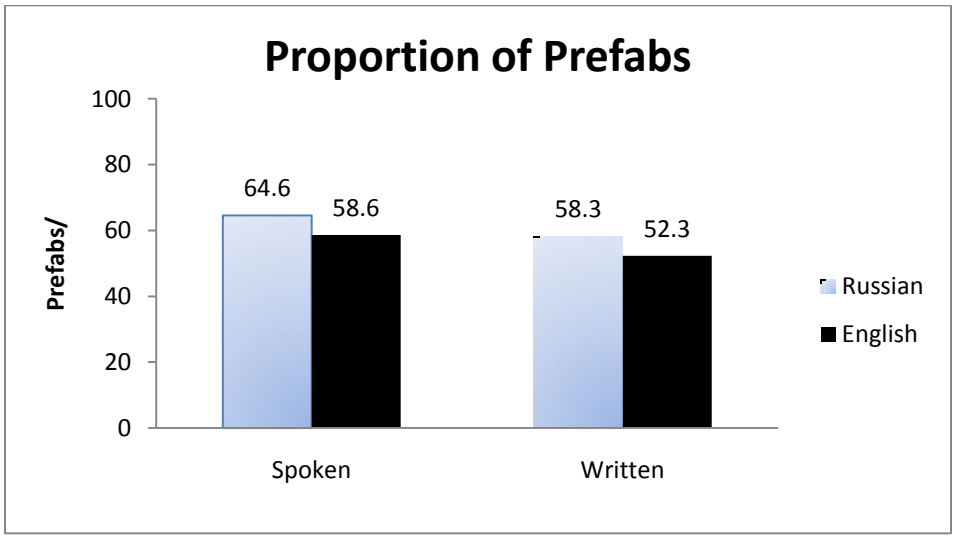

Figure 1: Proportion of prefabs in Russian and English texts

This study confirmed that speakers of a free word order language such as Russian rely on the idiom principle as much as speakers of a fixed word order language. In fact, the Russian data revealed on average 6% greater number of prefabs in texts. The

difference in the proportion of prefabs between the Russian and English data is explained by several factors.

First, to have more prefabs in Russian means the speakers stored and processed more chunks of language. Prefabricated sequences are processed holistically as units of language. *Chunking* as a cognitive mechanism increases the quantity of material that can be processed. The data suggests the Russian speakers process more of the language holistically than English speakers. Holistic processing minimizes the internal encoding and decoding work for the speaker and frees him to attend to other tasks in communication (Pawley and Syder, 1983). In Russian, the freed cognitive space is used for processing morphological inflections and syntactic manipulations associated with a free word order. The advantage of the holistic system is that it reduces processing effort. It is more efficient and effective to retrieve a prefabricated string than to create a novel one. On the other hand, the advantage of the analytic system, which creates grammatical strings out of small units by rule, is its flexibility for novel expression and the interpretation of novel and unexpected input. The trade-off exists between storage units and processing effort; holistically stored strings of multiple words require less cognitive effort in processing. An additional amount of analytic processing is required in Russian to process inflectional morphology and word order manipulations. This, in turn, is compensated by the increased number of prefabs and the increased length of prefabs.

Second, the expansion of the Russian corpus by 40,000 tokens and the addition of more genres to the text database make it possible only for the general comparison between English and Russian. Nevertheless, the results of the current study revealed that the speakers of a free word language such as Russian rely on the idiom principle even

more than the speakers of English. In other words, regardless of the corpora's size or genre, more than fifty percent of any Russian text is formulaic. Despite the differences between Russian and English, both languages exhibit a similar tendency: more than half of any text consists of prefabricated, multiword, holistically stored material. This finding strengthens the hypothesis that the idiom principle is a universal feature of the global organization of the linguistic system rather than a language-specific English phenomenon.

5.1.2 Distribution of prefab types in Russian and English

The comparison of the results of the distribution of prefab types in Russian and English reveals several distinctions, summarized in Table 2.

Table 2: Distribution of Prefab types in Russian and English

	Lexical		Grammatical		Pragmatic	
	Russian	English	Russian	English	Russian	English
Spoken	43.9%	38.8%	23.9%	20.5%	32.2%	16.7%
Written	73.1%	71.5%	17.6%	16.9%	9.3%	2.4%

The lexical group constitutes the largest number of prefabs for both Russian and English, followed by the grammatical and pragmatic groups. There is greater number of all types of prefabs in Russian than in English. (The group of reducibles is not analyzed in Russian; this group increases the overall number of prefabs in English, not included in this data.)

The difference between English and Russian is not statistically significant. The striking difference, however, is between spoken and written medium in both languages.

The spoken Russian has 43.9% and English, 38.8% of lexical prefabs while the written Russian has 73.1% and English, 71.5 % of lexical prefabs. Both languages exhibit a similar tendency where the number of lexical prefabs is greater in writing than in speaking. The difference between the two languages in both spoken and written media are not statistically significant.

In the grammatical group, there are more grammatical prefabs in spoken Russian than in spoken English: 23.9% of prefabs for spoken Russian and 20.5% for spoken English. In writing, Russian similarly has a greater number of grammatical prefabs, 17.6%, while English has 16.9%. The differences are not statistically significant.

The pragmatic group of prefabs exhibits the most differences. There are almost twice as many pragmatic prefabs in Russian as in English. This could be due to the genres of texts selected for the analysis, the participants, or the structural differences in two languages. Spoken Russian has 32.2% and spoken English only 16.7%, (significant, $p < 0.001$); written Russian has 9.3 % and written English, 2.4%, (significant, $p < 0.001$). Pragmatic prefabs in English occur almost exclusively in speech. Russian has the same tendency for the distribution of pragmatic prefabs, except the number of pragmatic prefabs overall is higher in Russian than in English. The difference is attributed to the text types and a greater variety of genres examined in the Russian study. The differences in the distribution of lexical, grammatical, and pragmatic prefabs in Russian and English are illustrated in Figure 2.

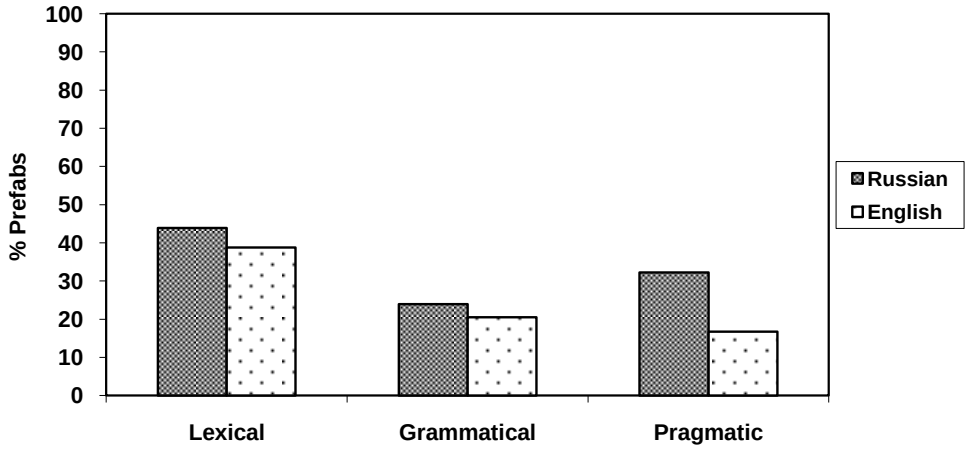

Figure 2: Distribution of prefab types in spoken Russian and English corpora

The differences in the distribution of prefab types between the two languages -- Russian and English -- are less significant than between the spoken and written medium within each language. The medium appears to have a greater impact on the distribution of prefabs than the language type, as illustrated in Figures 2 and 3.

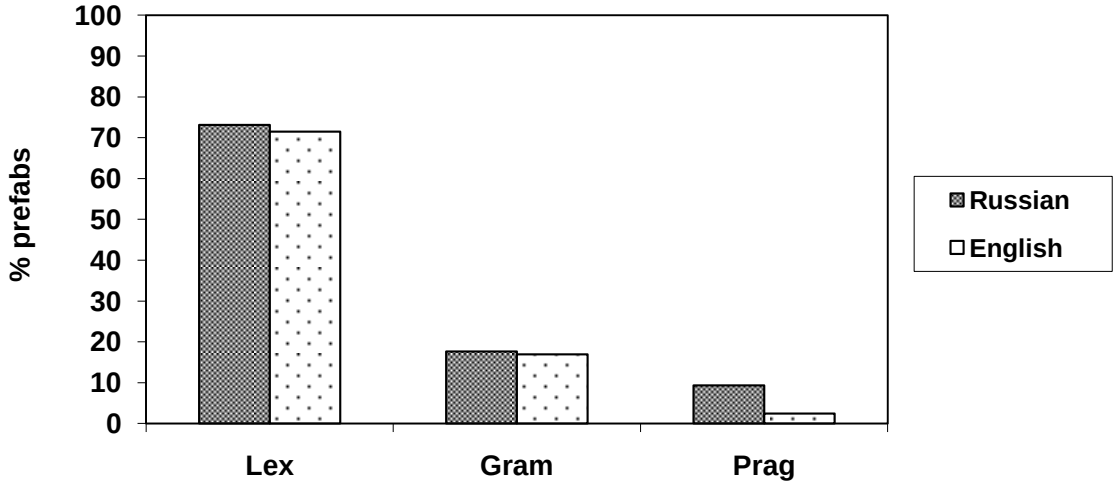

Figure 3: Distribution of prefab types in written Russian and English corpora

5.1.3 Length of Prefabs in Russian and English

Overall, the Russian prefabs are longer than the English ones, as shown in Table 3. The typical Russian prefabs are three to six words in length. The longest prefabs that were found in the corpus occurred in fairy tale texts and had up to16 words. Such length of prefabs is rare. The more typical ones are two to five words long. The frequency of a prefab seems to correlate with the length of a prefab; the longer the prefab, the less frequent it is. The average length of the spoken Russian prefab is 3.25 words while the spoken English prefab is only 2.61. I found, not surprisingly, that both Russian and English prefabs are longer in written language than in speech, which is due to the lack of pressures associated with online communication. And again the Russian written prefabs are longer than English written prefabs, 4.12 and 2.80 words/prefab, respectively. The length of the Russian prefabs can be attributed to the greater collocational frequency of elements in a clause and for much greater preference for a periphrastic coding rather than derivational for many concepts (Bybee et al., 1994). Table 3 illustrates the average length of prefabs in Russian and English; the comparison shows that the Russian prefabs are longer than the English ones.

Table 3: Average length of prefabs in Russian and English

	Russian	English
Spoken	3.25 word/prefab	2.61 word/prefab
Written	4.12 word/prefab	2.80 word/prefab

The difference in the length of prefabs between Russian and English is due to the fact that many Russian categories are expressed periphrastically.

Now let us consider the average length of prefab types—lexical, grammatical, and pragmatic in Russian and English. Table 4 summarizes the measures of length of prefabs in both languages.

Table 4: Average length of prefab types in Russian and English

	Russian	English
Lexical	4.03	3.03
Grammatical	2.98	2.26
Pragmatic	3.89	2.29

The table shows that lexical prefabs are the longest in both Russian and English. Russian lexical prefabs are the longest, averaging 4.03 words/prefab, while the length of English lexical prefabs is 3.03 words/prefab. The grammatical class is 2.98 for Russian and 2.26 for English prefabs. Grammatical prefabs are longer in Russian than in English due to the inflectional morphology and preference for periphrastic type of expression in Russian. Pragmatic prefabs are 3.89 words/prefab in Russian and 2.29 in English. The reason pragmatic prefabs are longer than grammatical prefabs is that many pragmatic prefabs include speech formulas that are used for everyday standard situations and are often a clause length: *How do you do?; I am sorry to keep you waiting; That's another story; I wouldn't dream of it;* etc. The Russian examples include: *Kak dela? How are things?; Chto novogo?(*What's new?*); Skol'ko let, skol'ko zim; (*How many summers, how many winters? i.e., I haven't seen you for ages*).*

Another measure considered in the study was the length of non-prefabricated expressions. Table 5 illustrates the length of both prefabs and non-prefabs in Russian and English.

Table 5: Average length of non-prefab and prefab strings in Russian and English

	NON-PREFAB		PREFAB	
	Russian	English	Russian	English
Spoken	2.89/2.43	2.47/2/16	3.25/4.29	2.61/4.01
Written	4.13/3.01	3.84/3.02	4.12/4.30	2.80/4.08

Because prefabs and non-prefabs represent different mechanisms of processing, the length of these expressions in conjunction with the measure of choices gives an idea of the processing effort involved in each type of mechanism. Non-prefabricated strings represent analytic processing, which requires assembling words one at a time. Analytic processing requires more cognitive effort in constructing language, which is reflected in a longer processing time. The time and attention demands associated with the spoken medium favors holistic processing. This leads to not only more prefabs in speaking than in writing in both languages, but also to the fact that both prefabs and non prefabs are shorter in speaking than in writing. A prefabricated "chunking" method --rather than constructing strings word by word according to a rule -- is what aids fluency and speed in online communication. Thus, non-prefabricated strings are also longer in written texts than in spoken. The written medium does not have the same time and attention pressures as the spoken medium. The writer has more time to formulate, rewrite and edit his or her message. This allows a writer more freedom in composing a language. As a result, the

non-prefabricated strings are longer in writing than in speaking. The limitations of online processing affect the chunking mechanism by shortening the number of words per prefab and non-prefab. These limitations of time pressure, focus and attention are not present in offline processing. More time allows more choices in coding the language.

The first figure under *non-prefabs* refers to non-prefab strings indicated by dashes in the analysis. The second figure refers to non-prefab strings, including also adjacent open slots and open slots occurring within prefabs, which tend to be short, thus explaining the lower average figure. The figures in Table 5 show that both prefab and non-prefab strings are longer in written than in spoken language. This is expected in light of the discussed differences associated with online and offline processing constraints.

5.2 Choice

A choice represents a cognitive effort made by a language user in encoding and decoding a message. The greater the number of choices made in constructing a language-- the greater is the cognitive effort involved in language processing. Traditionally, in compositional generative models, it is assumed that each position in a clause offers a choice. What the current study shows, however, is that the presence of prefabs in a clause reduces the number of choices a speaker must make in constructing a language. Retrieving each word separately from the mental storage represents a greater cognitive effort than choosing a pre-constructed, ready-made sequence of multiple words. The number of overall choices in Russian and English are compared in Section 5.2.1. The lexical choices are compared in Section 5.2.2. Discussion or the differences of spoken and written medium in Russian are presented in Section 5.3.

5.2.1 Analysis of choices in Russian and English

All modalities, text types, and genres contain prefabs, which constitute single multiword retrievals from our mental lexicon. Because of prefabs, the number of retrievals is fewer than the number of words in a text of any size. However, determining the number of prefabs in a text does not automatically reveal the number of retrievals, due to some variability inherent in prefabs. A separate analysis of choices was designed to reveal the number of choices in a text. This analysis of choices is indicative of the processing effort involved in composing a text, and thus, it allows the comparison between text types, genres, speakers, and languages.

The method is designed in such a way that only the choices of slot-fillers/words are considered. Choice of tense, mood, number for the verb, nominative form for the subject, or accusative form for the object are not considered as a matter of policy. In other words, in this analysis, one slot can never represent more than one choice. Thus, tense-forming auxiliaries do not represent separate choices. Table 6 summarizes the number of choices in Russian and English texts.

Table 6: Number of choices in spoken and written Russian and English

	Russian	English
Spoken	65%	68%
Written	78%	75%

Overall, 65% of all slots represent a choice in spoken Russian and 68% in spoken English, which is not statistically significant ($p > 0.001$). More choices are made in written than in spoken texts in both languages. In writing, we observe the following: 78%

of all slots represent a choice in written Russian and 75% in written English, which again is non-significant (3% difference, $p > 0.001$). The difference in the number of choices between writing and speaking in Russian is 13%, ($p < 0.001$), and in English is 7%, ($p < 0.001$), which is statistically significant in both cases. The observed difference in the number of choices is greater between the written and spoken medium in both languages rather than between the two languages.

5.2.2 Analysis of lexical choices in Russian and English

The lexical choices were considered separately because they represent a greater cognitive effort in processing than grammatical choices. It is supposed that choosing an item from a restricted set is less demanding than selecting one from an infinitely large set. Grammatical elements such as articles, prepositions, and auxiliaries represent a limited list, whereas lexical elements in comparison come from a large unlimited set. Therefore, the calculation of the lexical choices is conducted separately in addition to the overall number of choices; this is meant to give an additional estimate to the cognitive effort involved in processing. Table 7 shows two measurements: First, the number of slots filled by the lexical choices out of all slots in Russian and English texts is presented; second, the number of lexical choices out of the total number of choices made in Russian and English texts is shown.

Table 7: Number of lexical choices in spoken and written Russian and English

	Russian	English
Spoken	29% of all slots 41% of all choices	27% of all slots 39% of all choices
Written	46% of all slots 58% of all choices	41% of all slots 55% of all choices

Overall, more lexical choices are made in Russian than in English. More slots out of the entire number of slots available in texts are filled by lexical choices in Russian than in English. The data shows that 29% of Russian vs. 27% of English slots represent lexical choices. Thus, the difference, however, is not statistically significant when the two languages are compared.

The statistically significant difference in the number of lexical choices is between speaking and writing in both languages: the spoken Russian, 29%, vs. written Russian, 46%, of all slots; ($p < 0.001$); and spoken English, 27%, vs. written English, 41%; ($p < 0.001$). Thus, the greater difference in the number of slots filled by the lexical prefabs is observed between spoken and written media than between the two languages. The comparison of the Russian and English lexical slots shows 3% difference, which is statistically not significant.

The comparison of the number of lexical choices to the overall number of choices leads to similar conclusions. The difference in the number of choices between Russian and English is not statistically significant, whereas the difference between the spoken and written media is more pronounced. Thus, there is 41% of lexical choices of the overall number of choice in spoken Russian compared to spoken English 39%. There is 58% of

lexical choices of all choices in written Russian compared to 55% in written English. The variation in the number of lexical choices shows that the differences observed are more pronounced between spoken and written media rather than between the two languages. Both Russian and English demonstrate the same tendency in the number of lexical choices out of the total number of choices made in constructing a text.

5.3 Spoken and Written Russian Language: Genre Differences

It has always been clear that neither a spoken language nor written language is a unified phenomenon. Each mode itself allows a multiplicity of styles and genres. There may be an overlap between speaking and writing, in the sense that some kinds of spoken language may be very written-like, and some kinds of written language very spoken-like. That there are differences between the use of prefabs in spoken and written language, and among the different genres within each category, is not surprising. Biber (1986) points out that there are three important parameters of variation in speech and writing: 1) interactive vs. edited text, 2) abstract vs. situated content, and 3) reported vs. immediate style. The genres examined in this study vary across all three parameters, and the results indicate that prefabs are used most where the language is most interactive, most situated in content, and most immediate in style.

With the development of new technology, new forms of communication such as e-mail, texting, and chat come forth only to further blur the distinction between spoken and written language. A number of factors are responsible for differences in the kinds of language a person may use.

Aspects of written style may be borrowed by speakers when it suits them, just as aspects of spoken style may be borrowed by writers. The context of language use, the

purpose of the speaker or writer, the subject matter of what is being said or written – these are some of the factors that influence the form language takes. In regards to the use of prefabs, the distinction between spoken and written media demonstrated the strongest effect.

I found, not surprisingly, that the six genres of language analyzed in this study – two spoken and four written – were different from each other in the proportion, distribution, and length of prefabs. Some of the differences appeared to be caused only by the fact that the language was spoken or written. Others were more characteristic of the more formulaic genres such as fairy tales. More often there were additional factors of language use that interacted with the spoken and written distinction.

5.3.1 Spoken Russian corpus

Drawing on studies of English natural conversations, Pawley and Syder, 1983; Sinclair, 1995; Wray, 2002; Biber 1986, 1981, have argued that fluent and idiomatic control of a language rests to a considerable extent on the knowledge of multiword formulaic sequences. These formulaic sequences are often a clause length and what Pawley and Syder (1983:23) call "lexicalized sentence stems." These researchers define a lexicalized sentence stem as "a unit of clause length or longer whose grammatical form and lexical content is wholly or largely fixed" (ibid: 24). It has been noticed that many such stems have a grammar that is unique and is subject to an idiosyncratic range of syntactic restrictions. Many Russian prefabs are "frozen" structures and do not allow regular syntactic manipulation usually permitted in this free word order language. For example, the following Russian prefabs do not allow any changes in the word order; any

word order changes in examples a) through e) lead to the loss of grammaticality or idiomaticity:

a) sovershenno verno; V poslednee vremja
completely true *Lately*

b) Krome togo; Vse taki
Besides *Really*

c) Delo v tom chto; Mezhdu prochem
The thing is that *By the way*

d) S odnoi storoni …. s drugoi storoni…..
On the one hand….. *On the other hand….*

e) Tem ne menee; Tak vot
Nevertheless *Here we go*

The comparison of two genres of spoken Russian revealed a number of distinctions summarized in Table 8.

Table 8: Proportion of prefabs in spoken Russian

Genres	Slots	Slots filled by prefabs	% of prefab slots
Interviews	8,098	5,409	66.8%
Conversations	40,371	25,901	64.2%
	48,469	31, 311	64.6%

The interviews have a greater number of slots filled by portions of prefabs, 66.8%, than conversations, 64.2%. The 2.6 % difference is statistically not significant ($p > 0.001$). The greater number of prefabs in interviews was attributed to interviews

being more preplanned than natural conversations, to the individual skills of the participants, as well as to the setting and time pressures associated with professional interviews. Only a general comparison with the English spoken data is possible; there is a lack of genre distinction in spoken corpus in the English-based study.

Next, the distribution of prefabs types in two genres of spoken Russian corpus are shown in Table 9.

Table 9: Distribution of prefab types in spoken corpus

Genres	Lexical	Grammatical	Pragmatic
Interviews	2,288 (42.3%)	1,363(25.2%)	1,757 (32.5%)
Conversations	9,609 (37.1%)	6,060 (23.4%)	10,230 (39.5%)
	11,897 (37.9%)	7,427 (23.9%)	11,987 (38.2%)

The number of lexical prefabs is high in any text because they directly partake in the propositional content of an utterance. Overall, there are more grammatical prefabs in Russian than in English because of the inflectional morphology in a free word order language. The high percentage of pragmatic prefabs in both interviews and conversations is expected. Pragmatic prefabs occur almost exclusively in spoken medium because most of them have functions that could be indicated by punctuation, paragraphing, or in other graphic ways in written texts.

Table 10: Average length of prefabs in spoken corpus (number of words per prefab)

	Lexical	Grammatical	Pragmatic
Interviews	3.51	2.89	3.59
Conversation	3. 55	2.91	3.31

The difference in length of prefabs in interviews and conversation is statistically non significant ($p > 0.001$). The variation is minimal between the two genres of spoken medium.

5.3.2 Written Russian corpus

The written medium is not homogeneous. It comprises a variety of genres and styles used for various purposes. The written Russian corpus used for the study consists of four distinct genres—fiction (FC), popular-science writing (PSW), socio-political journalism (SPJ), and fairy tales (FT). Table 11 demonstrates 1) the overall number of slots represented by each genre; 2) the number of slots filled by portions of prefabs out of the total number of slots; 3) the percentage of slots filled by portions of prefabs in the four written genres of Russian.

Table 11: Proportion of prefabs in written Russian texts

Genres	Slots	Slots filled by prefabs	% of prefab slots
FC	5,214	3,071	58.9%
PSW	5,106	2,818	55.2%
SPJ	5,202	2,830	54.4%
FT	5,018	3,254	64.9%
	20,540	11,973	58.3%

The range in the number of slots filled with portions of prefabs across written genres analyzed is from 54.4% to 64.9%. The average proportion of slots filled with parts of prefabs in written language is 58%. The highest density of prefabs is in fairy tales (FT), at almost 65%. In this study, one fifth of the fairy tales corpus is composed of prefabs wholly repeated from one place to another, which is a typical characteristic of the

genre. The second genre with the highest percentage of prefabs is Fiction (FC), which contains 58.9%. Popular-science writing (PSW) and socio-political journalism (SPJ) follow with 55.2% and 54.4%, respectively. The difference between FT and SPJ is statistically significant ($p < 0.001$), as evidenced through chi-square goodness of fit testing.

Next, consider the distribution of prefab types across four written genres of Russian, presented in Table 12.

Table 12: Distribution of prefab types in written corpus

Genres	Lexical	Grammatical	Pragmatic
FC	2,254/ 73.4%	556/ 18.1%	61/ 8.5%
PSW	2,004/ 71.1%	486/ 17.3%	328/ 11.6%
SPJ	2,063/ 72.9 %	504/ 17.8%	263/ 9.3%
FT	2,434/ 74.8%	556/ 17.1%	264/ 8.1%
	8,755/ 73.1%	2,102/ 17.6%	1,116/ 9.35%

The greatest number of prefabs in written texts is lexical, at 73.1%, followed by grammatical prefabs, 17.6%, and pragmatic, 9. 35%. The range for lexical prefabs across various genres of written text is 3.75%, with the highest number for fairy tales, 74.8%, and then in fiction, 73.4%. The lowest frequency is in PSW, 71.1%, followed by SPJ at 72.9%.

The second group of prefabs in terms of frequency of occurrence in writing is the grammatical prefabs. The range for grammatical prefabs across written genres is 1%, which is not significant ($p > 0.001$). The highest number of grammatical prefabs is in

fiction at 18.1 %, and the lowest is in Fairy tales at 17.1%. In the middle of the range are PSW at 17.3% and SPJ at 17.8%.

Pragmatic prefabs are not frequent in writing in comparison with speech; the range across written genres is 3.5%. The highest number of pragmatic prefabs is in SPW at 11.6%, and the lowest is in fairy tales at 8.1%. Fiction is at 8.5% and SPJ at 9.3%. The lexical prefabs are most prominent in writing, with grammatical prefabs next and pragmatic being the third group in terms of frequency.

If we compare the distribution of prefab types between spoken and written media, the differences are striking. The number of lexical prefabs in writing is almost twice as high as in speaking, 73.1% vs. 37.9%. The average proportion of grammatical prefabs in speech is 23.9% and in writing, 17.6%, for a difference of 6.3 percent. The number of pragmatic prefabs is significantly higher in speech, 38.2%, than in writing, 9.4%, with the difference at 28.8%. The data clearly show that the distribution of prefab types is significantly greater along the speech-writing continuum, rather than along speaker and genre variation.

Table 13: The average length of prefabs in written Russian corpus (words per prefab.)

	Lexical	Grammatical	Pragmatic
FC	4.21	2.8	3.9
PSW	3.81	2.7	3.7
SPJ	3.78	2.9	3.8
FT	4.19	2.9	4.3

Prefabs tend to be longer in writing than in speech; non-prefabricated units are also longer in writing than in speech. Results are summarized in Table 10. The average length of prefabs in writing is 4.12 words/prefab. The range is from 4.21 lexical prefabs in FC to 2.7 grammatical prefabs in SPW. The longest prefabs are lexical; the average length of lexical prefabs is 4 words/prefab. In FC, we see 4.21 words/prefab; in FT, 4.19; PSW at 3.81; and SPJ at 3.78 word/prefab.

Pragmatic prefabs are the second group in length measure. The average length of pragmatic prefabs is 3.9 words/prefab. Variation includes the longest prefabs, 4.3 words/prefab in fairy tales, and close measures in the other three genres: - 3.9 word/prefab in fiction; 3.7 in PSW; and 3.8 in SPJ.

Grammatical prefabs form the third group. The average length is 2.83 words/prefab with 2.8 words/prefab in fiction, 2.7 in SPW, and 2.9 in both SPJ and fairy tales. Tables 10 and 11 reveal that the length of prefabs contrasts between speech and writing to a greater extent than across genres or speakers.

5.4 The Role of Word Order and Case Marking in Formulaic Language

The notion of word order flexibility traditionally attributed to a type of language such as Russian states that all logically possible six types of word orders are perceived by native speakers to be grammatical and acceptable. However, the current study of Russian prefabs revealed that prefabs in Russian are fixed conventional structures that exhibit a lack of flexibility in word order. Prefab components have a fixed position within a prefab, and any syntactic transformation results either in a loss of their meaning, grammaticality or prefab status.

Prefabs in a free word order expedite language processing. They facilitate the speed and fluency of language processing, particularly in spoken medium. Extra encoding and decoding is required for processing inflectional information and word dependencies in a free word order language. Prefabs, however, minimize the processing work involved. They serve the needs of a speaker as well as a hearer in language processing. Word order variation is not random, boundless, and unpredictable. First, because all languages are put to similar uses in the human communities in which they are spoken, it is common to find that speakers of a language use distinct word order patterns for such universally relevant discourse tasks as 1) introduction of participants into narrative, and 2) disagreement with claims of another speaker, etc.

Second, because users of all human languages share common cognitive capacities, the word order variations permitted in any language must conform to the constraints imposed by those capacities. Alternative word orders serve different discourse functions. Possible motivation for the correlations observed is the flexibility of the word order in Russian. Variation seems to be restricted along the prefab boundary lines. In other words, Russian prefabs exhibit the same level of fixity as English prefabs. Prefabs cannot be broken like non-prefabs for syntactic manipulation. However, mapping of sounds to concepts is conventionalized, even in a free word order language.

5.5 Conclusions

5.5.1 The social functions of prefabs

Several recent models intrinsically accommodate some or all aspects of formulaicity, including Cognitive Grammar (Langacker, 1987, 1990, 1991, 1999); Construction Grammar (Fillmore, Kay & O'Connor, 1988; Michaelis & Lambrecht,

1996; Tomasellso and Brooks, 1999); the Emergent Lexicon (Bybee, 1998); Radical Construction Grammar (Croft, 2001); Lexical-Functional Grammar (Bresnan, 1982a, 1982b); and Pattern Grammar (Hunston and Francis, 2000).

Many scholars have pointed out that language is about communication: Its primary function is phatic and subjective, rather than propositional, and centered on the connection between speaker and addressee (Jakobson, 1960, Thompson and Hopper, 2001, Scheibman, 2002). In speaking of the study of poetics within linguistics, Jakobson (1960) points out that all languages have different means for accomplishing different functions: "No doubt, for any speech community, for any speaker, there exists a unity of language, but this overall code represents a system of interconnected subcodes; each language encompasses several concurrent patterns which are characterized by a different function." Formulaic language is one strategy that languages have to accomplish the simultaneous goals of efficiency of expression and ease of processing, in addition to the important aim of building of solidarity between a speaker and a hearer by using familiar phrasing to express a specific message. Prefabs have, in this sense, an important social indexicality (Agha, 2007). One of Jakobson's main points is to emphasize the inseparability of linguistics from poetics and the importance of including poetic language within the domain of linguistic study. He speaks of one of the properties of poetry as its ability to produce the "conversion of a message into an enduring thing", to reify a message through its reiteration. This is true of the development of prefabs as well and is an important factor in the social function of this linguistic strategy. Agha (2003) suggests that "particular styles and forms of language allow cultural values to be maintained and communicated across social populations", and formulaic expressions are an important

reflections on how particular, value-laden, messages can be repeated, and how new messages can be naturalized through the use of formulaic frames or more schematic prefabs.

Results of the current study suggest that a language with free word order contains on average 64.6% in speaking and 58.3% in writing of prefabs in written and spoken discourse. A greater number of grammatical prefabs in a free word order language is due to coding of grammatical case and other notions present in this morpho-syntactic type of language. Speakers of Russian use the idiom principle as much as English speakers do, thus alternating between the use of the idiom principle as well as the open-choice principle in language production. Russian data strengthens the hypothesis that the idiom and the open-choice principles are not language-specific but are general principles of language production grounded in human cognition and conceptualization, which allow the minimizing of processing costs and increase in fluency of the speakers.

In the future, the next step would be to test the method and hypotheses on more typologically different languages such as agglutinative, analytic, synthetic and polysynthetic. This method may not be applicable to the languages where word boundaries are difficult or impossible to identify. While the notion of "word" may be more problematic for some language than others, additional cross-linguistic studies of typologically different languages are necessary to test these hypotheses and strengthen this line of argumentation.

References

Agha, Asif. 2003. The social life of cultural value. *Language and Communication*, *23*:3-4, 231-73

------. 2007. *Language and Social Relations*. Cambridge University Press.

Altenberg, B. 1998. On the phraseology of spoken English: The evidence of recurrent word-combinations. In A.P. Cowie (ed.), *Phraseology: Theory, analysis, and application*, 101-22. Oxford: Clarendon Press.

Biber, D. 1986. Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language, 62*:2, 384: 414.

Biber, D. 1991. *Variation across speech and writing*. Cambridge University Press.

Bod, R. 2001. Examplar-based syntax: How to get productivity from examples. *The Linguistic Review, 23*, 291-320.

Bussmann, H. 1996. *Routledge Dictionary of Language and Linguistics*. London and New York: Routledge.

Butler, C. S. 1997. Repeated word combinations in spoken and written text: Some implications for functional grammar. In C.S. Butler, J. Connolly, R. Gatward, & R. Vismans (eds.), *A fund of ideas: recent developments in functional grammar, pp.* 60-77. Amsterdam: University of Amsterdam.

Bybee, J. 2007. *Frequency of use and organization of language*. Oxford: Oxford University Press.

Bybee, J. 2006. From usage to grammar: The mind's response to repetition. *Language*, *82*, 711-733.

Bybee, J.L. 2002. Phonological Evidence for Examplar Storage of Multiword Sequences, *SSLA, 24*, 215-221.

Bybee, J.L. 2001 Sequentiality as the basis of constituent structure. In *Evolution of Language out of Pre-language*, T.Givon & B.F. Malle (eds), 109-134. John Benjamins Publishing Company: Amsterdam/Philadelphia.

Bybee, J. 1998. The emergent lexicon. CLS 34: *The Panels*, 421-35. University of Chicago: Chicago Linguistic Society.

Bybee, J. & McClelland, J. 2005. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *Linguistic Review, 22*, 381-410.

Bybee, J., Perkins, R. & Pagliuca, W. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: The University of Chicago Press.

Bybee, J. & Torress, R. 2009. The role of prefabs in grammaticization: How the particular and the general interact in language change. In R. Corrigan, E. Moravcsik, H. Ouali, K. Wheatley (eds.) *Formulaic Language, Vol. II,* John Benjamins Publishing Company: Amsterdam/Philadelphia.

Cacciari, C. & Tabossi, P. 1988. The comprehension of idioms. *Journal of Memory and Language, 2*: 668-683.

Cacciari, C. & Tabossi, P. (eds.) 1993. *Idioms: Processing, Structure, and Interpretation.* Hillsdale, New Jersey: Lawrence Erlbaum.

Carnie, A. 1994. *Syntax: A Generative Introduction*. Blackwell Publisher.

Chafe, 1994. *Discourse, consciousness and time: the flow and displacement of conscious experience in speaking and writing.* Chicago: University of Chicago Press.

Chomsky, N. 1957. *Syntactic structures*. The Hague: Mouton.

------. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

------. 2006. On phrases. In R. Freidin, C. Otero and M. Zubizaretta (eds.) *Foundational issues in linguistic theory*, 133-66. Cambridge, MA: MIT Press.

Chomsky, N. & Halle M. 1968. *The sound pattern of English*. New York: Harper & Row.

Conklin, K. & Schmitt, N. 2008. Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29*, 72-89.

Corbett, G., Hippisley, A., Brown, D. and Marriott, P. 2001. Frequency, regularity and the paradigm: A perspective from Russian on a complex relation. In *Frequency and the Emergence of Linguistic Structure,* J.L.Bybee and P.Hopper (eds), 201-229. John Benjamins Publishing Company Amsterdam/Philadelphia.

Croft, W. 2001. *Radical construction grammar: Syntactic theory in typological perspective.* Oxford: Oxford University Press.

Cronk, B. & Schweigert, W. 1992. The comprehension of idioms. *The effects of familiarity, literalness and usage, 13*(3): 131-146.

DeKeyser, R. M. 1997. Beyond Explicit Rule Learning. Automatizing Second Language Morphosyntax. *SSLA, 19*, 195-221.

Dryer, M 1983. Coos Word Order. Paper presented at the Western Conference on Linguistics. University of Oregon, Eugene.

Dryer, M. 1988. Object-Verb Order and Adjective – Noun Order: Dispelling a Myth. *Lingua, 74*, 77-109.

Ellis, N. 1996a. Sequencing in SLA . Phonological Memory, Chunking and Points of Order. *SSLA, 18*, 91-126.

Ellis, N. 1996b. Analyzing Language Sequence in the Sequence of Language Acquisition. *SSLA, 18*, 361-368.

Ellis, N 1997. Morphology and Longer Distance Dependencies. Laboratory Research Illuminating the A in SLA. *SSLA, 19*, 145-171

Erman, B. & Warren B. 2000. The Idiom Principle and the Open Choice Principle. *Text 20*, 29-62.

Everaert, M., Van der Linden, E.J., Schenk, A. & Schreuder (eds.). 1995. *Idioms: Structural and Psychological Perspectives*. New Jersey: Lawrence Erlbaum.

Fenk-Oczlon, G. 2001. Familiarity, information flow, and linguistic form. In *Frequency and the Emergence of Linguistic Structure,* J.L.Bybee and P.Hopper (eds), 431-448. John Benjamins Publishing Company Amsterdam/Philadelphia.

Frazier, L., Clifton, C., and Randall, J. 1983. Filling gaps: Decision principles and structure in sentence comprehension. *Cognition*, *13*, 187-222.

Garcia, E. & Florimon von Putte. 1989. Forms are silver, nothing is gold. *Folia Linguistica Historica, 8*(1-2): 365-84.

Gibbs, R.W. 1980. Spilling the beans on understanding and memory for idioms in conversation. *Memory and Cognition*, *8*(2): 149-56.

Gibbs, R.W. & Nayak, N. 1989. Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology, 21*(1): 100-38.

Gibbs, R.W., Nayak, N., Bolton, J., & Keppel, M. 1989. Speakers' assumptions about the lexical flexibility of idioms. *Memory and Cognition, 17*(1): 58-68.

Goldberg, A. 1995. *Constructions: A Construction Grammar Approach to Argument Structure.* Chicago: The University of Chicago Press.

Goldberg, A. 2006. *Constructions at Work: The Nature of Generalization in Language.* Oxford: Oxford University Press.

Greenberg, J. 1963. Some Universals of Grammar with Particular Reference to the Ordering of Meaningful Elements. *Universals of Language* ed. by Joseph Greenberg, 58-90. Cambridge, Mass: MIT Press.

Gregory, M.L., W.D. Raymond, A. Bell, E. Fosler, Lussier, and D. Jurafsky. 1999. The Effects of collocational strength and contextual predictability in lexical production. *CSL, 35*.

Hale, K. 1992. Basic word order in two "free word order" languages. In Doris Payne (ed.), *Pragmatics of word order flexibility*. John Benjamins, 63-82.

Hay, Jennifer. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics, 39*(6): 1041-1070.

Hawkins, J. 1983. *Word Order Universals*. New York: Academic Press.

Hopper, P 1998. Emergent Grammar. In *The New Psychology of Language*. M. Tomasello (ed.) Lawrence Erlbaum Associates, Publishers.

Hunston, S. &Francis, G. 2000. *Pattern Grammar: a corpus driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Hymes, D.H 1962/1968. The ethnography of speaking. In T. Gladwin & W.C. Sturtevant (eds.) *Anthropology and Human Behavior*. Washington, DC: Anthropological Society of Washington, 13-53.

Irujo, S. 1986. A piece of cake: learning and teaching idioms. *ELT Journal, 40*(3): 236-242.

------. 1993. Steering clear: avoidance in the production of idioms. *International Review of Applied Linguistics in Language Teaching, 31*(3):205-219.

Israel, Michael. 1996. The way constructions grow. In A. E. Goldberg (ed.), *Conceptual structure, discourse, and language*, 217-30. Stanford, CA: CSLI.

Jackendoff, R. 1997. *The architecture of the language faculty*. Cambridge, MA: MIT Press.

Jakobson, R. 1957 [1971]. Shifters, verbal categories and the Russian verb. Reprinted in *Roman Jacobson, Selected Writings II*, 130-47, The Hague: Mouton.

------. 1960. Closing statement: Linguistics and poetics. In Thomas Sebeok (ed.), *Style in Language*. Cambridge, Mass.: The MIT Press.

------. 1966 [1971]. Quest for the essence of language. Diogenes, 51. Reprinted in *Roman Jacobson, Selected Writings II*, 345-59. The Hague: Mouton.

------. 1990. Some questions of meaning. In L.R. Waugh (ed.), *On language: Roman Jakobson*, 315-23. Cambridge, MA: Harvard University Press.

Jurafsky, Daniel, Alan Bell, Michelle Gregory, & William D. Raymond. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In Joan Bybee & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 229-253. Amsterdam: John Benjamins.

Langacker, R. 1987. *Foundations of Cognitive Grammar. Vol. 1: Theoretical* . Stanford, CA: Stanford University Press.

------. 1990. *Concept, image, and symbol: The cognitive basis of grammar* [Cognitive Linguistic Research 1]. Berlin & New York: Mouton De Gruyter.

------. 1999. *Grammar and conceptualization* [Cognitive Linguistic Research 14]. Berlin & New York: Mouton de Gruyter.

------. 2000. A dynamic usage-based model. In M. Barlow and S. Kemmer (eds.), *Usage-based models of language*, 1-64. Stanford, CA: CSLI Publications.

------. 2008. *Cognitive Grammar A basic introduction*. Oxford: Oxford University Press.

Lehmann, W. 1973. A Structural Principle of Language and its Implications. *Language, 49*, 47-66.

Li, Charles N. 1975. *Word order and word order change. Austin*, TX: University of Texas Press.

Li, P., Elizabeth Bates and Brian MacWhinney. 1993. Processing a language without inflections: a reaction time study of sentence interpretation in Chinese. *Journal of Memory and Language, 32*: 169-92.

Lonngren, L. 1993. Chastotnyj slovar' sovremennogo russkogo jazyka*. (Acta Universitatis Upsaliensis, Studia Slavica Upsaliensis 33). University of Uppsala: Uppsala.*

MacWhinney, B. 1997. Implicit and Explicit Processes. Commentary. *SSLA, 19*, 227-281.

MacWhinney, B. 2001. Emergentist approaches to language. In *Frequency and the Emergence of Linguistic Structure*, J.L. Bybee and P. Hopper (eds), 449-469, John Benjamins Publishing Company: Amsterdam/Philadelphia

Major, R.C. 1996. Chunking and Phonological Memory. A response to Ellis. *SSLA, 18*, 351-354.

Maier, I. 1994. Review of Lennart Lonngren (ed.) Castotnyj slovar' sovremennogo russkogo jazyka. *Rusistika Segodnja, 1,* 130-6.

Mel'cuk, I. 1988. Dependency syntax: theory and practice. Albany: State University of New York Press.

Mel'cuk, I. 1998. *Collocations and lexical functions*. In A. P. Cowie Phrasiology: theory, analysis and applications. Oxford: Clarendon, 23-53.

Michaelis, L.A. & Lambrecht, K. 1996. Toward a construction-based theory of language function: the case of nominal exraposition. *Language, 72*(2): 215-247.

Miller, G.A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2): 81-92.

Mithun, M. 1992. Is basic word order universal? In Doris Payne (ed.), *Pragmatics of word order flexibility.* John Benjamins, 15-62.

Moon, R. 1997. Vocabulary connections: multi-word items in English. In Schmitt & McCarthy (eds.) *Vocabulary: description, acquisition and pedagogy*. Cambridge: Cambridge University Press.

Moon, R. 1988a. Fixed expressions and idioms in English. Oxford: Clarendon Press.

Moon, R. 1988b. Frequencies and forms of phrasal lexemes in English. In A.P. Cowie(ed.) *Phraseology*. Oxford: Clarendon Press, 79-100.

Nattinger, J. & DeCarrico, J (1993). Lexical phrases and strategic interaction. *Georgetown University Roundtable on Language and Linguistics.* Georgetown: Georgetown University.

Nichols, J. 1986. Head-Marking and Dependent-Marking Grammar. *Language, 62,* 56-119.

Pawley, A. and Syder, F.H. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In *Language and Communication*, J.C. Richards and R.W. Schmid (eds), 191-226. Longman: London.

Payne, D. 1990. *The Pragmatics of Word Order: Typological Dimensions of Verb Initial Languages.* Berlin: Mouton de Gruyter.

Payne D. 1992. ed. *Pragmatics of Word Order Flexibility*. Amsterdam: John Benjamins Ruthlidge Dictionary of Language and Linguistics.

Perkins, M.R. 1999. Productivity and formulaicity in language development. In M. Garman, C. Letts, B. Richards, C.Schelletter & S.Edwards. *Issues in normal & disordered child language: from phonology to narrative.* Special Issue of The New Bulmershe Papers. Reading: University Reading, 51-67.

Reali, Florencia, & Morten Christiansen. (2006). Word chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology, 60*(2): 161-170.

Schmitt, N. 2004. *Formulaic Sequences*. Amsterdam and Philadelphia: John Benjamins.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (ed.), *Formulaic sequences,* 127-151. Amsterdam and Philadelphia: John Benjamins.

Scheibman, J. 2002. Point of view and Grammar. Amsterdam and Philadelphia: John Benjamins.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Swinney, D.A., and Cutler, A. 1979. The access and processing of idiomatic expressions. *Journal of Berbal Learning and Verbal Behavior*, *22*, 358-374.

Stubbs, M. 1996. *Text and corpus analysis.* Oxford: Blackwell.

Tanenhaus, M., & Carlson, G. (eds.). 1989. *Linguistic structure in language processing.* Dordrecht; Boston: Kluwer Academic Publishers.

Thompson, S. and Hopper, P. 2001. Transitivity, clause structure, and argument structure: Evidence from conversation. In *Frequency and the Emergence of Linguistic Structure,* J.L.Bybee and P.Hopper (eds.), John Benjamins Publishing Company Amsterdam/Philadelphia.

Tomasello, M. & Brooks. P.J. 1999 Early syntactic development: a construction grammar approach. In M. Barrett (ed.), *The development of language*. Psychology Press, 161-190.

Tremblay, Antoine, Bruce Derwing, and Gary Libben. 2007 April 18-21. *Are lexical bundles stored and processed as single units?* Paper presented at UWM Linguistics Symposium on Formulaic Language, Milwaukee, WI.

Thompson, S. 1978. Modern English from a Typological Point of View: Some Implications of the Function of Word Order: *Linguistiche Berichte, 54,* 19-35.

Vennemann, T. & Harlow, R. 1977. Categorial Grammar and Consisten Basic VX Serialization. *Theoretical Linguistics, 4,* 227-254.

Vogel Sosa, A., and MacFarlane, J. 2002. Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*. *Brain and Language, 83*: 227-236.

Vos, S., Gunter, H., Kolk, T.C., Herman, H.J., & Gijsbertus, M. 2001. Working memory constraints on syntactic processing: An electrophysiological investigation. *Psychophysiology, 38*, 41-63.

Weinert, R. 1995. The Role of Formulaic Language in second language acquisition: A Review. *Applied Linguistics, 16*(2): 180-205.

Wong Fillmore, L. 1976. *The second time around: cognitive and social strategies in second language acquisition*. Unpublished PhD thesis, Stanford University.

Wong Fillmore, L. 1979. Individual differences in second language acquisition. In C.J. Fillmore, D.Kempler & S-Y.W.Wang (eds.) In *Individual differences, language ability and language behavior*. New York: Academic Press, 203-239.

Wood, M.M. 1986. *A definition of idiom*. Bloomington, IN: Indiana University Linguistics Club.

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Wray, A. 2008. *Formulaic Language: Pushing the Boundaries*. Oxford: Oxford University Press.

Wray, A. & Perkins, M.R. 2000. The functions of formulaic language: an integrated model. *Language and Communication, 20*(1): 1-28.

Zuidema, W. 2006. What are the productive units in natural language grammar? A DOP Approach to the automatic identification of constructions. *Proceedings CoNLL 2006*, 29-36.