Master's thesis presented to the Faculty of Arts of the University of Zurich for the degree of Master of Arts UZH

Institute of Computational Linguistics

Words in Mind and Words in Language

Author: Dolores Batinić
Student ID: 09-735-184

Examiner: Prof. Dr. Barbara Sonnenhauser
Supervisor: Dr. Tanja Samardžić
Submission date: 10.12.2015

Abstract

In recent years, many psychologists as well as computational linguists have investigated the overlap between human associations, namely, words which are recalled as response to a given stimulus, and textual co-occurrences. However, the content of both data sets as well as of their overlap has not been studied in detail. In the present work, I investigate the relation between Russian associations and textual co-occurrences on bases of lexical and semantic characteristics of the two. The findings of this work reveal that the overlap of semantic and lexical properties between associations and co-occurrences depends on the part-of-speech of the target words, as well as of their antonym potential. The results suggest that the correlation between associations and co-occurrences may not be language-independent.

Acknowledgements

I want to express my sincere thanks to my supervisors Barbara Sonnenhauser and Tanja Samardžić for their encouragement and support. I am also very grateful to Mirjam Zumstein for her very appreciated technical assistance and constant availability, and to Noah Bubenhofer for offering me his extremely helpful suggestions. I also want to express my best thanks to my brother Josip for proofreading this and many other papers I wrote during my studies, and to my friend Maja Škrkić for commenting parts of this work. Above all, I would like to thank Michael Greeff for many long and inspiring discussions, most valuable advice and moral support. Last but not least, I want to thank my parents for their love and continuous encouragement.

Contents

1 Introduction

An association or associative response is the first word that comes to mind after having perceived another word - the associative stimulus. People tend to associate *young* to *old*, *black* to *dark* and *snow* to *white*. Associations have been of interest in the field of psychology since the end of the 19th century, when they were believed to represent a window into human unconsciousness (Jones 1964). Over the years they have become a tool for investigating the semantic net and mechanisms underlying semantic memory (Steyvers *et al.* 2004). Recently, associations have gained interest of computational linguists investigating corpus-based extraction of semantically related words (Schulte im Walde & Melinger 2008). In both fields, the research of associations has been based on the assumption that they reflect words occurring in common context, phenomenon commonly known as the principle of contiguity:

> "Objects once experienced together tend to become associated in the imagination, so that when any one of them is thought of, the others are likely to be thought of also, in the same order of sequence or coexistence as before" (James 1890:561)

Relying on the principle of contiguity, many researches measured the overlap between associations and words that occur together in language, often referred to as co-occurrences (Church & Hanks 1990; Wettler *et al.* 2005). However, the linguistic properties of the overlap between associative pairs and co-occurrences have been overlooked; although it has already been observed that co-occurrences cover paradigmatic as well as syntagmatic types of associations (Rapp 2002; Wettler *et al.* 2005; Sahlgren 2006), not much effort has been done in order to quantitatively corroborate these observations.

In order to investigate the commonalities and the differences between associations and co-occurrences within this work, I will attempt a novel approach by measuring not only the overlap on a word level, but also in parts-of-speech and in semantic relations. The focus of my research will be Russian associations, gathered in Ufimceva *et al.* (2004). Analysing Russian associations is motivated by the assumption that Slavic associations display a different distribution of semantic relations than those traditionally described in research, originating mostly from English and German-speaking area (Anstatt 2008). The prediction of Russian associations with corpus-based methods has recently been subject of research (Panchenko *et al.* 2015). However, in the task described in Panchenko *et al.* (2015), associations have not been

used for investigating the contiguity hypothesis, but rather for evaluating corpus-based word similarity measures. Moreover, the semantic properties of associations and co-occurrences, as well as the reasons for successful or failed predictions were not further discussed.

Besides investigating the contiguity hypothesis, with this work I am hoping to contribute to a better understanding of human associations, their semantic and lexical properties as well as the extent to which they can be described as being language-independent.

In the first section of this thesis, I will explain what association norms are, how they are gathered, how semantic relations within associations have been described in research and under which assumptions and with which success have they already been predicted with corpus-based methods. In the second section I will concentrate on the corpus material and association norms I used for this investigation. The focus of the third section will be the analysis of Russian associations and their comparison with associations given by speakers of other Slavic languages. A significant part of the section will be dedicated to the methodology I used to define and analyse semantic relations within associations. In the fourth section the co-occurrences extracted from Russian corpus will be compared to human associations. The focus of the section will be the analysis of the lexical overlap, as well as of the overlap between linguistic characteristics of associations and co-occurrences. By the end of the chapter the challenges of using association norms and corpus-based methods for investigating the contiguity hypotheses will be discussed.

2 Theoretical background

In the following sections I will explain what associations and association norms are, what do they consist of and how have they been used in research. I will particularly focus on explaining the ideas behind association measures as well as measures of semantic similarity and the link between these measures and human associations. I will also dedicate a section for explaining the types of semantic relations and highlighting various hypotheses which have been proposed in order to describe the constants within association norms.

2.1 Association norms

In order to compare co-occurrences with human associations, collections of associations commonly known as association norms are used. Association norms contain a set of cue words (stimuli), to which the participants in an associative experiment are invited to recall the first word that comes into their mind. In most cases, only one response is asked and the participants are given a time limit to complete the task. At the end of the experiment, the frequencies for each response are counted. Usually, there is one word that is commonly recalled by a high number of participants. The most frequent response is otherwise known as the primary response. The rest of the responses are distributed over a high number of participants and tend to be more individual. For instance, in association norms analysed by Raible (1981:10), in 75% (752 out of 1008) of participants recalled the same response to the stimulus *slow*, whereas the responses ranked as second, third or fourth most common were given by only 2% of the participants each. This type of distribution is commonly known as power-law distribution (De Deyne & Storms 2008). Associations are otherwise referred to as word associations, semantic associations, free recall and free word associations.

The study of word association norms has a long tradition in psychology. The first association test was developed by Kent & Rosanoff (1910) following the work of Francis Galton (1880). The participants were instructed to say or write the first word they recall after having read a cue word from the list of 100 stimuli. The research was conducted on 1000 participants of varied educational backgrounds and professions with equal participation of both genres. The collected and quantified responses are known as first association norms. Although the Kent & Rosanoff list of stimuli has certain shortcomings (it consists almost exclusively of nouns and adjectives) it has been frequently used as the basis for several other association tests in the 60s

and 70s, which were gathered, for instance, for Germanic and Romance languages (Postman & Keppel 1970), as well as for Polish (Kursz 1967). Besides the Kent & Rosanoff list, another commonly used association test was written by Palermo and Jenkins (1964) and consists of 200 stimuli of various parts-of-speech.

The research on associations was continued with the Edinburgh Association Thesaurus (EAT; Kiss *et al.* 1973), in which each of the 100 participants was presented a list of 100 randomised stimuli out of a total of 8400 cue words. The first step towards a large-scale collection of association norms was presented by Nelson *et al.* (2004) with the collection of association norms of the University of South Florida. The collection was gathered for more than 20 years with the purpose of building the largest association database collected in the United States. The association norms of the University of South Florida were build with the data coming from more than 6000 participants who were asked to produce a response to the set of 5019 stimuli.

Within English speaking areas, not much is known about the collections of association norms gathered in Slavic speaking countries, where the interest towards association norms has increased starting from the 70s. The first collection of Russian association norms was collected during 1969 and 1970 by Lenot'ev (1977). The list of 500 stimuli was presented to Russian speaking participants aged 16 to 50 years and having different education levels. The number of responses varied from 200 to 600, depending on the stimulus. The Leont'evs collection of Russian association norms was followed by the association norms collected by Karaulov *et al.* (1994ff) from 1988 to 1997 and by Russian-French association experiments conducted from 1998 to 2000 (Filippovič *et al.* 2002). One of the most recent collections of Russian association norms, comprising of also Belarusian, Bulgarian and Ukrainian language and containing 112 (almost) identical stimuli for each of the four languages was collected from 1998 to 1999 in Moscow, Kursk, Minsk, Sofia and Nizhyn by Ufimceva and collegues (2004) and it has been known as Slavic Association Dictionary.

In line with the *offline* associative experiments, a new form of gathering association norms for Russian language has been realized in an online platform named Sociation.org[1]. Sociation.org is an experimental crowdsourcing project with the aim to build the largest collection of association norms for Russian language. The experiment has been running since 2009. Currently, Sociation.org consists of 349'812 responses to 38'537 stimuli. A major dif-

[1]http://sociation.org/ (18.10.2015).

ference between Sociation.org and other association experiments mentioned above consists in the game-effect of the project - the participants are not only invited to write their response to a fixed list of stimuli as in the traditional associative experiments, but they can also suggest new stimuli themselves, to which other participants are invited to respond. Depending on the number of added stimuli and responses, the rating of the participant goes up.

The association norms mentioned above have been used in numerous studies in different areas of research and for different purposes. Initially, the association tests were conducted with the aim of defining *normal* reaction to certain stimuli and detecting divergences in responses of mentally unstable patients (Jung 1918). Over the years, the use of association norms has changed track - instead of being seen as an expression of subconsciousness, they have become a valuable resource for exploring the structure of mental lexicon and semantic memory. The collections of association norms have revolutionized the view on the structure of semantic networks, which has previously been based on the assumption that words in mind are connected by similarity (Collins & Quillian 1969) or by unstructured, arbitrary graphs (Collins & Loftus 1975). According Steyvers & Tenenbaum's (2004) most recent developments in statistical analysis of association norms at the University of South Florida, the structure of semantic networks builds upon a few words, presumably acquired in early age and frequent in language, which become gradually more connected to a high number of many other words. For example, some words which were observed to be connected to many other words are *food, money, water, car* and *good* (ibid.:54). Moreover, it has been shown that any two words in a semantic net built from association norms consisting of more than 5000 words are separated by a maximum of 5 associative steps (ibid.).

Imagining association norms as an insight into human mind has been criticised by the linguist Jane Aitchison. First, because thinking up an immediate response to just one word can be seen an "unnatural kind of activity"; second because responses may be altered depending on the previously mentioned words (priming effect); and third, because the participants are invited to give only the first answer that comes into their minds (Aitchison 2013:100). From another point of view, it may as well be argued that although responding to a list of stimuli might represent an unnatural kind of activity, the associations represent the most natural connection between words, since they are the most intuitive reactions to stimulus words (Sahlgren 2006).

In the last few decades, association norms have been used for exploring the correlation

between words people tend to associate and words which often co-occur (Church & Hanks 1990; Rapp 2002; Lemaire & Denhière 2006; Schulte im Walde 2006). The task of extracting these co-occurrences which may overlap with human associations has also been referred to as "association prediction" (Rapp & Wettler 1991) or as "computation of word associations" (Rapp & Wettler 1993). When predicting associations with corpus-based methods, the computational methods attempt to produce links between words in the corpus that correspond to the links between associative pairs.

2.2 Corpus-based methods for predicting associations

Corpus-based methods that have been used for prediction of associations can be divided into methods of first-order and second-order statistics. In the field of computational linguistic, the methods of first-order statistics are often referred to as *association measures* or *methods for co-occurrence extraction*. The methods of second-order statistics are often referred to as *similarity measures*. The main purpose of both is to extract words from textual corpora that are connected to each other in a particular way, either by common occurrence in text or by being "somehow related"[2].

Association measures have often been used in order to extract collocations (Evert 2005).[3] Similarity measures have been implemented for computational applications such as word and document clustering (Pereira *et al.* 1993; Huang 2008), analyses of topics of a document (Griffiths & Steyvers 2003), paraphrasing (Fernando & Stevenson 2008), sentiment analysis (Turney & Littman 2003), etc.

By extracting co-occurrences from text corpora, association measures have been regarded as being able to provide robust estimates of traditionally gathered word association norms, the latter being "costly and unreliable". (Church & Hanks 1990:22). The underlying assumption was that people associate words that are frequently used together in language.

[2]The notion of semantic similarity in distributional semantics does not refer only to synonyms, but to all the words which might be though of being somehow semantically related: "although *buy* is semantically very similar to *acquire* on the one hand, and *car*, *vehicle* and *van* are all somehow semantically related, the notion of "semantic similarity" actually covers different types of semantic relations, each with different logical properties and each licensing different inferences" (Lenci 2008:20-21).

[3]The term *collocation* was coined in the 1930s by J. R. Firth to describe characteristic, or habitual word combinations. Whereas the term co-occurrence may refer to words co-occurring in a particular range of text, such as a sentence, paragraph, document or another pre-define context window, *collocation* refers exclusively to two words that frequently appear in immediate proximity. For instance, *strong coffee* is a collocation, whereas *powerful coffee* is not, since the words*powerful* and *coffee* are rarely expressed together by native speakers of English.

Similarity measures have also been regarded as being able to provide estimates of human associations: the performance of similarity methods has often been evaluated either by using synonym dictionaries, human similarity judgements, TOEFL tests or by comparing the retrieved similar words with human association norms (Rapp 2002). The underlying assumption was that associations primarily consist of similar words. Hence, without knowing what association norms exactly consist of, they have been employed in the evaluation of both association measures as well as of similarity measures.

Whereas in the field of artificial intelligence, computational linguistics, and information retrieval both association and similarity measures are aimed toward fulfilling a particular task, in psychological research of the semantic net the aim is to find a model that would represent the way human memorise and recall semantic information. In consequence, other than being just a tool for word retrieval, association measures as well as similarity measures have been hypothesised to represent the way human mind works. In the following sections I will explain what these models consist of and how successfully they have been employed in the task of predicting human associations.

2.2.1 Association measures

Church & Hanks (1990) proposed the *association ratio* as a method for extracting co-occurrences based on the notion of mutual information, which, according to their study, can estimate human associations directly from computer readable corpora. They reported that the association ratio measured with mutual information score captures co-occurrences similar to the responses from association norms from Palermo and Jenkins (1964). For instance, they found that the extracted co-occurrences of *doctor* such as *dentists, nurses, treating, treat, and hospitals*, etc. (1990:23) tend to overlap with the most common associative responses to the cue word *doctor*, such as *nurse, sick, health, medicine, hospital, man, sickness.*

The mutual information compares the probability of observing two units u and v together (joint probability) with the probability of observing u and v independently (chance). According to Fano (1961), the measure of mutual information can be explained as follows: if u and v have probabilities P(u) and P(v), then their mutual information, I(u,v), is defined as:

$$I(u,v) \equiv \log_2 \frac{P(u,v)}{P(u)P(v)} \tag{1}$$

The formula for mutual information can be computed as follows: for a particular candidate pair *u, v* where *u* denotes the cue word and *v* the candidate co-occurrence, a contingency table (cfr. Table 2.1) is computed, in which the values for variables *a,b,c* and *d* are stored. The variable *a* denotes the number of times *u* and *v* appear in the same context; *b* denotes the number of times *u* and any word other than *v* appear in the same context; *c* denotes the number of times *v* and any word other than *u* appear in the same context, and *d* the number of times that any two words other than *u* and *v* appear in the same context.

Table 2.1: Contingency table (the sign ¬ is to be read as *anything but x*)

	v	¬v
u	a	b
¬u	c	d

According to the contingency table, given N as the sum of all pairs in the contingency table, the values for mutual information score can be calculated as follows:

$$MI(u,v) = \log_2 \frac{aN}{(a+b)(a+c)} \tag{2}$$

Church & Hanks (1990) estimated word probabilities by counting the number of observations of *u* and *v* in three different corpora, having sizes from 15 million to 36 million words. They calculated joint probabilities by counting the appearances of *u* followed by *v* in a window of *w* words and normalizing by the size of the corpus. They stated that by using smaller window sizes, the fixed expressions (such as the idiom *bread* and *butter*) are expected to be retrieved, whereas when larger window sizes are chosen, mutual information is expected to capture semantically related concepts, such as *man* and *woman*. For their calculations Church & Hanks used the context window of ±5 words (5 words before and after the cue word), which they estimated to be a compromise for being able to capture the advantages of considering both small and large context windows. Church & Hanks did not propose a systematic evaluation measuring to what extend the predicted associations resemble those of human association norms.

The principle of mutual information was used by Rapp & Wettler (1991) for investigating the overlap between co-occurrences and English associations norms by Russel & Jenkins (Jenkins, 1970). According to their results, there is a correlation between predicted responses

(co-occurrences) and human responses since the ranks of the predicted responses are much higher than they would have been by chance (Rapp & Wettler 1991). They measured the correlation by defining their own task-specific method, namely by calculating "the mean rank of the 5 most frequent human responses when looked up in the simulation word list, but weighted by the frequency of the responses and averaged over all 100 association examples" (ibid.:27). In the same study, Rapp & Wettler compared different types of context windows: sentences windows, which include all words in a sentence and continuous windows, which contain a constant number of words. They reported the optimal context window to be around 18 words and noted no particular difference between the results of continuous and sentence-based context windows. Rapp & Wettler (1993) repeated the experiments related to prediction of German and English human associations collected by Russell & Jenkins (Jenkins 1970) by using the corpora collections of 33 million words for English and 21 million words for German. They evaluated the results by considering the average primary responses for average subjects and they found that with the context window comprising ± 12 words, in 17% of the cases for English and 19% for German the predicted response was equal to the human primary response (Rapp & Wettler 1993). On average, the predicted primary response was given by 12.6% of the English subjects and 6.9% of the German subjects (ibid.). In order to evaluate the proportion of the overlap, they compared it to the average of 37 primary responses given by an average English subject and 22.5 responses given by an average German subject.[4] Rapp & Wettler (1993) noted that, although there was an expected bias toward syntagmatic associations (associative pairs having words of different parts-of-speech), many paradigmatic associations (associative pairs having words of equal part-of-speech) were correctly predicted as well, such as *man-woman, black-white, bitter-sweet*, etc. (ibid.:92).

Rapp (2002) investigated the prediction of 100 association norms from Edinburgh Associative Thesaurus (EAT) collected by Kiss *et al.* (1973) by using log likelihood, another measure widely used for co-occurrence extraction (Dunning 1993; Daille 1994; Orliac & Dillinger 2003; Lü & Zhou 2004). Referring to the contingency table in Table 2.1, and with the assumption of the binomial distribution, the log likelihood can be determined by contrasting the likelihood

[4]That means that the overlap between most frequent co-occurrences and most frequent human responses was compared to the number of times that an average person had recalled exactly that response, which, at the end of the experiment, turned out to be the most frequent response for the given stimulus. The comparison of the overlap with the proportion of participants recalling the primary response has also been performed in other studies (cfr. Wettler *et al.* 2005).

of observing the counts in the contingency under the hypothesis of independence ($L(H_0)$) with the likelihood of these counts under the hypothesis of dependence ($L(H_1)$) and considering the logarithm of the result (Seretan 2010:41):

$$LLR(u, v) = -2 \log \frac{L(H_0)}{L(H_1)} \tag{3}$$

Considering the values from the contingency table, the log likelihood can be computed as follows:

$$\begin{aligned} LLR = 2(&a \log a + b \log b + c \log c + d \log d \\ &-(a+b) \log(a+b) - (a+c) \log(a+c) \\ &-(b+d) \log(b+d) - (c+d) \log(c+d) \\ &+(a+b+c+d) \log(a+b+c+d)) \end{aligned} \tag{4}$$

The primary co-occurrence, i.e., the word with the highest log likelihood score (v), is considered to appear more frequently with the cue word (u) than it is to be expected by chance. In the context of association prediction, it corresponds to the predicted primary response (v) to a given stimulus (u).

For predicting associations with log likelihood, Rapp (2002) applied context window of ±20 words to the texts from the British National Corpus. The predicted primary response was equal to the observed primary response in 27% of the cases, meaning that the predicted response to 27 out of the 100 stimuli (i.e. the extracted co-occurrences with the highest log likelihood score) was equal to the observed primary response. Rapp compared the overlap of 27% to the average of 28 primary responses given by an average subject (ibid.).

Perhaps the best degree of overlap between co-occurrences and human associations was shown by Wettler *et al.* (2005), who calculated the co-occurrence statistics for 100'106'008 words from the British National Corpus and compared the results with associations from the Edinburgh Associative Thesaurus, which were restricted to 100 words used originally by Kent and Rosanoff (1910). By applying the formula from associationist learning theories (Schwartz & Reisberg 1991) and considering the context of ±10 words, Wettler and colleagues (2005) found that 29 out of 100 predicted primary responses corresponded to human primary

responses, whereas in comparison 28 responses of an average participant coincided with the primary response (ibid.). Moreover, they found that 64 of the 100 predicted primary responses have been produced by at least one participant (ibid.).

An in-depth look into the co-occurrence distribution and its relation with human associations has been provided by Schulte im Walde & Melinger (2008). They based their investigation on responses to 330 verb stimuli that were collected in web experiment, in which native German speakers were asked to write as many associations for one stimulus as they could in a time span of 30 seconds. Schulte im Walde & Melinger analysed the effect of the context window on the coverage of associations within co-occurrences. They confirmed the observations of Church & Hanks (1990) that co-occurrences found within small context window sizes tend to be related by some syntactic function whereas words that co-occur within larger context window sizes larger might point "towards situational or even world knowledge" (ibid.:107). Contrary to what has been hypothesised by cognitive psychologists (Clark 1970), Schulte im Walde & Melinger observed that no semantic relation such as antonymy or synonymy dominated in German association norms, but that they mostly reflected other types of semantic relations, such as those in *drizzle–wet*, *munch–yummy*, *defrost–water*, etc. (2008:119). The exact proportion of semantic relations within human associations and co-occurrences was, however, not quantified. Although they based their investigation exclusively on verb stimuli, Schulte im Walde & Melinger hypothesised that the insights obtained from their study could be generalised to other parts-of-speech and other languages as well (ibid.).

As shown above, various studies have measured the overlap between co-occurrences and human associations. The best results predicted slightly less than a third of human associations. The context window varied from ± 5 to ± 20 context words, whereas the best performance has been achieved by the context window of ± 10 words by Wettler *et al.* (2005). However, in the reported studies no mention has been made about the possible influence of the choice of stimuli words on the overlapping statistics nor about the reasons why one model performs better than another.

2.2.2 Similarity measures

Similarity measures are based on the assumption of distributional hypothesis, stating that words with a similar meaning occur in similar contexts (Harris 1954). The more common

co-occurrences the two words have, the more they are believed to be related. Some of most recent models for retrieving semantically related words include Latent Semantic Analysis (LSA; Deerwester *et al.* 1990; Landauer & Dumais 1997) and topic model (Griffiths & Steyvers 2003).[5] LSA is presented as being able to simulate "several psycholinguistic phenomena", (Landauer & Dumais 1997:211), since the idea behind word retrieval with LSA is to group together two words that share similar contexts, as well as words that co-occur in the same passages of text (Landauer *et al.* 1998). The combination of co-occurrences and similar words seems to be promising concerning the prediction of human associations, since associations as well as the output of LSA consist of words which are connected to each other by different patterns (cfr. Griffiths & Steyvers 2003; Wandmacher *et al.* 2008). However, the critique of LSA has relied on the fact that it mostly extracts similar words and fails to retrieve the associations based on common occurrences of words (Griffiths & Steyvers 2002).

The topic model proposed by Griffiths & Steyvers (2003) seeks to eliminate the similarity bias of LSA by combining semantic similarity with frequency information ("[...] when a word is very diagnostic of a small number of topics, semantic context is used in prediction. Otherwise, word frequency plays a larger role", ibid:5). In the comparison made by Griffiths & Steyvers (2003), LSA correctly predicted primary response for 507 out of 4544 stimuli (11,16%) from association norms collected by Nelson *et al.* (2004), whereas the topic model correctly predicted 585 associative pairs (12,87%).

Another method that may be used to predict human associations is Hyperspace Analogue to Language (HAL), proposed by Lund & Burgess (1996). In HAL, a moving window of (usually) 10 words is passed across the corpus. With the shifting of the window, the co-occurrences are computed by regarding the distance of the target word to any other word within the window. The words in immediate proximity to the target word receive a higher arbitrary value (10), whereas if there are three or more words between the context and the target words, the assigned arbitrary value is lower (6). Hence, the HAL model saves the positional information, which plays a major role in simulating associations that correspond to co-occurrences and preserves the asymmetric character of human associations.[6] Although

[5]The co-occurrences for measuring LSA are collected from discrete passages of text. The context in this case is not defined as a context window of n-words around the target word, but it is extended to paragraph, article, chapter, document or even a whole book.

[6]"In HAL's high-dimensional space, tiger is the 4th neighbor to leopard, whereas leopard is the 1335th neighbor to tiger - an asymmetry in the direction one would find with word norms [...] China is Korea's 6th neighbor in HALs hyperspace; Korea is China's 40th neighbor" (Lund & Burgess 1996:15-16).

a correlation of HAL model with human associations was established by the authors (2000), they did not report the ratio of correctly predicted associative pairs.

The most recent computational models seeking to provide better representation of semantic memory are Bound Encoding of the Aggregate Language (BEAGLE; Jones & Mewhort 2007) and Recurrent Neural Network based Language Model (RNNLM; Mikolov *et al.* 2013a), a novel method containing two distinct neural models (CBOW and skip-gram), which is able to detect various linguistic regularities in word representations.[7]

In comparison to similarity measures, association measures seem to be more adaptable to the task of simulating human associations. Firstly, they support the traditional theory of associations according to which associations originate from contiguity in language (James 1890). Moreover, the association measures preserve the asymmetry observed between predicted associative pairs (for instance, the fact that the response *news* appears with a certain frequency to the stimulus *good* does not mean that *good* must appear with the same frequency to *news*. Similarly, the co-occurrence strength between *good-news* is not necessarily the same as between *news-good*), which is not the case for common similarity measures such as LSA. Moreover, it has been proven that similarity models (other than topic model) do not preserve the structure of associations, which is based on power-law distribution and displays the *small world* characteristics.[8]

A cognitive plausibility of both association measures and similarity measures being models of human semantic memory can be debated. The scepticism towards the distributional methods for predicting semantic association is due to the fact that the latter do not contain any extralinguistic information and the notion of context is reduced to the textual level. For instance, in order to understand that the word *nurse* is associated to female nurse and the word *engineer* to male engineer, extralinguistic knowledge is required. The same accounts to colours, shapes, and other matter that people perceive in spatio-temporal context. However, some effort is currently being put into incorporating the extralinguistic information into the

[7] A famous example of the capacity of the model to achieve linguistically motivated representation of word space is the recognition of gender information: for instance, if *king*, *man* and *woman* each represent a vector, the resulting vector for *king* - *man* + *woman* is situated very close to *queen* (Mikolov *et al.* 2013b:746)

[8] The challenges for corpus-based models as models of semantic memory have been discussed in Steyvers & Tennenbaum (2005), who have assumed that models which aim to simulate human semantic memory must produce the output having the similar structure as human associations, having short average path lengths between words, strong local clustering and following a power-law distribution (ibid.). They argue that these regularities have not been found within popular models of semantic structure including those based on high-dimensional vector spaces such as LSA, but that they may be found in topic model (ibid.).

methods of distributional semantics (Bruni *et al.* 2014).

2.3 Semantic relations within associations

As discussed in Sections 2.2.1 and 2.2.2, in one line of research, associations have been assumed to consist of similar words and were accordingly used as evaluation set for models extracting similar words. In the other line of research, associations have been assumed to reflect contiguity in text rather than similarity, and were accordingly compared to co-occurrences extracted by association measures. In neither case has there been a systematic investigation of the types of semantic relations within associative pairs. The fact that one can make different assumptions related to the nature of human associations implies that the type of relation holding together associative pairs is still under-investigated. However, there have been different attempts to explain the semantic constants within associations in both psychological and linguistic research.

Maybe the most common approach towards describing the type of relation within associative pairs is the division in paradigmatic and syntagmatic associations. The paradigmatic relation is said to appear within associative pairs in which stimulus and response display the same part-of-speech (for instance, black[ADJ] - dark[ADJ]). The syntagmatic relation, in contrary, is said to appear within associative pairs in which stimulus and response display different part-of-speech (for instance, black[ADJ] - coffee[NOUN]).

A long-standing observation in the association literature is that stimuli tend to elicit paradigmatic responses (Thumb & Marbe 1901; Deese 1962; Raible 1981). In the psychological research this statement has been extended with the remark that adults tend to produce prevalently paradigmatic associations whereas children tend to produce syntagmatic responses (Nelson 1977). This finding arose the discussion regarding the phenomenon of *syntagmatic-paradigmatic shift*, which has been estimated to occur between the age of 7 and 10 (ibid.:113). According to Brown & Berko (1960) (reported in Wettler *et al.* 2005:115), about 74% of the associative responses of adults were paradigmatic, whereas first grade children produced 72% syntagmatic responses.

In the line of research related to the *syntagmatic-paradigmatic shift* Brown and Berko (1960), Ervin (1961), Deese (1962), Cramer (1968) as well as Palermo (1971) found that noun stimuli tend to stimulate more paradigmatic responses than other word classes in all age

groups. That noun stimuli tend to attract paradigmatic responses was confirmed already in Entwisle's (1966) analysis of association norms of different age groups, in which she showed that kindergarten children as well as college students mostly associate noun responses to noun stimuli.[9] In contrary to noun stimuli, those belonging to other parts-of-speech showed a tendency to attract syntagmatic responses (Deese 1962; Cramer 1968; Nelson 1977). The traditional hypothesis that especially noun stimuli elicit mostly paradigmatic pairs has recently been argued by De Deyne & Storms (2008), who also noted that nouns constitute the most frequent part-of-speech of responses even if stimuli are adjectives or verbs. Paradigmatic responses have been observed more frequently by native speakers and syntagmatic responses by foreign language speakers (Meara 1982; Coulthard *et al.* 2000).

As noted already by Nelson (1977), instead of adopting the terminology *paradigmatic* or *syntagmatic* to describe relations between associative pairs, it would be more appropriate to use the division homogeneous/heterogeneous, the first referring to associative pairs which preserve the part-of-speech of the stimulus and the second referring to those which do not. Indeed, the structuralist dichotomy paradigmatic/syntagmatic (Saussure 1916) was not introduced in order to describe relations between associative pairs, but between linguistic entities, which can either be replaced or combined in order to produce a valid syntactic sequence. Although words that are in a paradigmatic relation usually have the same part-of-speech, it is not a necessary condition.[10] Similarly, part-of-speech heterogeneity between two words does not imply that there is a syntagmatic relation: as Nelson (1977:95) mentions, the adjective *hard* is a common response to the noun *difficulty*, but the two words are not syntactically combinable and hence they are not to be defined as syntagmatic but as heterogeneous.

Another opposition used for determining the relation between two terms is the distinction contiguity and similarity. Contiguity accounts for the terms which are expected to appear in the same context (same spatio-temporal context or field of experience), whereas similarity accounts for the terms semantically similar or opposite to one another. As Blank explains, since the opposites of particular unit may be expressed as much or less similar to that unit, the opposition may be seen as a kind of similarity as well (1997:143). The dichotomy *similarity* and *contiguity* has its roots in the antiquity: Aristotle explained the relations among ideas

[9]The number of participants was 1'600 for kindergarten children and college students (Entwisle 1966 in Nelson 1977:99).

[10]For example, a noun which can be replaced by a pronoun in the same sentence can be defined as being in a paradigmatic relationship with the respective pronoun, although they do not share the same part-of-speech.

in terms of contiguity, similarity and contrast.[11] In modern history, the discussion about contiguity and similarity as principles of thought processes has been developed by British empiricists Locke, Hume, Hartley and John Stuart Mill, who are considered predecessors to psychological theories of association (Blank 1997). The principles of contiguity and similarity were also emphasized by Gestalt psychology, which tries to define principles at the basis of human ability to acquire and maintain meaningful perceptions in an apparently chaotic world.

By relying on the principles of contiguity and similarity, Raible (1981) described associative pairs by outlining four types of relations: syntagmatic contiguity (white-snow), semantic contiguity (snow-winter), similarity (black-dark) and contrast (white-black). Under syntagmatic contiguity Raible described all relations between associative pairs belonging to different parts-of-speech such as *white-snow, red-blood,* and *music-sweet* (ibid.:19). As semantic contiguity he described relations between associations underlying the "proximity law", where proximity has been understood as temporal, spacial or textual, such as *tobacco-smoke* or *girl-beauty* (ibid.:24). As similarity Raible considered relations of synonymy, near-synonymy and categorical relations (hyponymy, hyperonymy, etc.), while contrast has been defined as antonymy in its widest sense: Raible's view of contrast is not restricted to antonym adjectives as it would be in a strict definition of antonyms (Lobanova 2012), but comprises a wide range of relations which can be though of as opposites in some way (gender oppositions like *boy/girl, he/she, king/queen;* adjectives of sensory perception like *black/white, dark/light, hard/soft, loud/soft, heavy/light*, situational opposites like *me/you*, as well as converse relations like *sell/buy* (Raible 1981:23). A relation between associative pairs which goes beyond semantic criteria is phonetic similarity, which according to Raible's observations appears to be very rare (ibid.). Raible based his study of human associations on 48 stimuli from American, German and French associations selected from Kent & Rosanoff list. Raible estimated that paradigmatic associations, i.e. contrast, similarity and semantic contiguity, are encountered in 80% to 90% primary responses.

According to Raible, the distribution of semantic relations is said to depend on whether the stimulus has an opposite (contrast) or not: if stimulus has an opposite, the response is expected to be the opposite; if stimulus does not have an opposite, the response is expected

[11] The roots of these terms can be traced back to Aristotle's De Memoria et Reminiscentia: "And this is exactly why we hunt for the successor, starting in our thoughts from the present or from something else, and from something similar or opposite, or neighbouring" (Sorabji 2004).

to stand in a relation of semantic contiguity with the stimulus. The remaining two relations, syntagmatic contiguity as well as similarity are, according to Raible, underrepresented in human associative pairs (ibid.).

As an example of predominance of the contrast rule, Raible reports that 83% of the participants (839 out of 1000) of an association experiment based on Kent & Rosanoff list associated *light* to *dark* and only 0.05 % (55 out of 1000) associated *night*, while even less participants recalled something else (ibid.:12).[12] Raible's observation about the primary contrast response is in line with psychological research conducted by Deese, who found that the strongest indication that two words are antonyms is that each is given as a response to another in association tests (Deese 1965 in Fellbaum 1998). According to Raible, if a word contains more opposites, then the sparsity of the responses is expected to be higher, like in the case of the stimulus *sweet*, which, having two opposites (bitter and sour), did not provoke such a high contrast reaction as the stimulus *dark* (43% participants responded *bitter* and 8% *sour*, ibid.). Raible assumed that associations are mostly paradigmatic and did not analyse the distribution of semantic relations for each part-of-speech separately, but for all chosen stimuli together.

A dominant role of contrast in associative norms was noted amongst others by Clark in *Word Associations and Linguistic Theory* (1970). Clark explained the contrast associations as being produced according to the *Minimal-Contrast rule*. By relying on the structuralists' assumption that the meaning of words can be explained by defining their distinctive set of features (man: +human, +male), Clark (1970) stated that the contrast is produced as a result of feature polarity alteration (woman: +human, -male).[13]

The relations between associative pairs collected for Germanic and Slavic languages have been studied by Tanja Anstatt (2008). According to Anstatt, Raible's hierarchy of associations does not account for Slavic languages. Inspired by Suprun (1983), who claimed that the tendency to paradigmatic responses is less pronounced in Slavic languages than in English, Anstatt investigated the distribution of semantic relations between 40 associative pairs (all

[12]Raible underpins the argument that the primary response tend to be in contrast with the observation that even healthy subjects very often coincidently express the exact opposite of what they intended (e.g. "Ihm war auch kein Berg zu *niedrig*" (No mountain was *low* enough for him, ibid., my emphasis).

[13]Clark assumes the existence of a pre-defined order of features, stating that the last feature is the one to be altered: "Consider the stimulus man. At stage one, comprehension entails setting up a list of features that completely characterizes this surface realization, perhaps as follows: [+Noun, +Det-, +Count, +Animate, +Human, +Adult, +Male].[...] At stage two, some associating rule is applied. If the rule were 'change the sign of the last feature', the associating mechanism would alter [+Male] to [-Male]. And then, at stage three, production would form the realization of the altered feature list [+Noun, +Det-, +Count, +Animate, +Human, +Adult, -Male] as woman." (1970:274).

having a potential opposite) for five languages mentioned above. To underpin her claim, Anstatt analysed the distribution of semantic relations within association norms for Germanic and Slavic languages collected in Russell (1970), Miller (1970), Kurcz (1967), Karaulov *et al.* (1994ff) and Leont'ev (1977).[14] Germanic languages she has taken into account are German and English, Slavic languages are Polish and Russian. Anstatt followed Raible's idea that there are four main semantic relations between associative pairs: similarity, contrast, and semantic/syntagmatic contiguity. She has also mentioned the existence of a *word-formation association*, introduced to describe associative pairs such as *dark-darkness*.[15] In contrast to Raible, Anstatt did not consider the categorical relations (hyponymy, hyperonymy, co-hyponymy) as a subset of similarity but as a separate category.

However, Anstatt's criteria for assigning semantic relations were generally in line to those of Raible.[16] According to Anstatt's results, Raibles hierarchy does not account for Slavic languages because the semantic relation which was found to be the most frequent in the primary responses of Russian and Polish association norms is not contrast, but syntagmatic contiguity. The presence of syntagmatic contiguity is observed to be particularly pronounced for Russian, where it accounted for almost three quarters (73%) of all analysed primary responses, as shown in Table 2.2 (2008:22).

Table 2.2: Proportion of syntagmatic contiguity between associative pairs from a total of 40 associative pairs having a contrast (Anstatt 2008:22)

	German	English	Polish	Russian
syntagmatic contiguity as primary response	8%	5%	50%	73%

Since Anstatt observed that the syntagmatic contiguity appeared to be the most frequent

[14]For a precise overview of meta-information concerning the related association tests and its participants see Table 1 in Anstatt (2008:13).

[15]Anstatt 2008:15: *Wortbildungsassoziation*, translation D.B.

[16]More precisely, as contrast she considered complementary antonymy (x or no x), scalar antonymy (more x or less x), vectorial antonymy (come–go) and converse (buy-sell); as similarity synonymy and near-synonymy; as semantic contiguity meronymy (finger-hand), words belonging to the same frame and subtypes of metonymy. However, Anstatt described syntagmatic contiguity as consisting of words that often co-occur and added that it is often produced by naming a prototypical example for one category - for instance, if the stimulus is an adjective, it would be in a syntagmatic relation with a "prototypical" noun which has been recalled (2008:15). Anstatt's definition of syntagmatic contiguity is arguable, since two terms can be though as being related through syntagmatic contiguity even if they are not prototypical of one another. For instance, when considering associative pair тихий-Дон (quiet-Don) (2008:20), it is arguable whether one can claim that being quiet (тихий) is a prototypical characteristic of the river Don, but the two words can produce a syntagma. Although the use of the notion of prototypicality in the definition of syntagmatic contiguity may be discussed, Anstatt's examples representing this category in principle mostly correspond to those mentioned by Raible.

relation among associative responses in Russian and Polish, in her study the contrast as primary response has come to be clearly underrepresented in relation to its distribution in German and English, as shown in Table 2.3.[17]

Table 2.3: Proportion of contrast relations between associative pairs from a total of 40 associative pairs having a contrast (Anstatt 2008:22)

	German	English	Polish	Russian
contrast as primary response	85%	75%	38%	23%

While, according to Raible, semantic contiguity is the second most represented semantic relation amongst associations, Anstatt reported that in Slavic associations semantic contiguity as well as similarity are only weakly represented (ibid.). The difference between Raible's and Anstatt's results is shown in Table 2.4, where the order of semantic relations between associative pairs is almost reversed.

Table 2.4: A language-independent hierarchy of associations according to Raible (1981) and hierarchy of Slavic associations according to Anstatt (2008)

	Raible	Anstatt
1.	contrast	syntactic contiguity
2.	semantic contiguity	contrast
3.	similarity/syntactic contiguity	similarity/semantic contiguity

The difference between Raible's and Anstatt's results may be attributed to various factors, first of which can be the time of experiments. Although Raible referred to Thumb & Marbe's experiments from 1901, his material for English, German and French collected from Postman & Keppel dates to 1970, whereas Anstatt's material dates from 1957 to 1990, each experiment being made in a different time-span (for German from 1957 to 1958, for English from 1961 to 1962, for Polish from 1964 to 1965, for Russian from 1988 to 1990, and from 1969 to 1970). Another reason for incompatibility of the results could be the fact that the number of participants in the associative experiments was not the same. Both analysed only a limited selection of associative pairs (48, respectively 40) and they did not quantify the presence of poorly represented semantic relations (similarity and syntagmatic, respectively, semantic

[17] Anstatt explained the lower frequency of contrast relation as a possible reason for an apparent "lower bundling" presented in Russian and Polish responses, in which at least more than a quarter of all participants recalled the same response (2008:23). In contrary, only 8% of Russian and 23% of Polish primary responses analysed by Anstatt were expressed by all participants (ibid.). However, Anstatt reported neither the total number nor the variance of the responses.

contiguity). Moreover, they did not consider the effects of the possible influence that the part-of-speech of the stimuli - which has been observed to play an important role by the response production in psychological research - could have exerted to the results of the analysis. Lastly, it remained unclear to what extent associations are language-independent and whether Slavic languages do indeed display a similar behaviour for what associations are concerned. It is therefore legitimate to question the distribution of semantic relations within associations as well as the peculiarities related to Slavic languages further in this work.

2.4 Previous research on predicting Russian associations norms

Predicting Russian association norms was one of the tasks at the first workshop on Russian Semantic Similarity Evaluation (RUSSE), which was held at the Dialogue[18] conference. The first task was to predict similar words such as synonyms, hypernyms or hyponyms (the dataset originated from human judgements of semantic relatedness between words). The second task was to predict human associations collected in Russian Association Dictionary and in Sociation.org. A total of 19 teams participated in the task, using a wide range of approaches (distributional models, decision trees, neural networks, etc.) and different lexical resources, such as Wikipedia, Wiktionary, linguistic ontologies, etc. (Panchenko *et al.* 2015). The methods were applied to corpora of different sizes and genres, such as Wikipedia, Russian National Corpus, web corpora, Twitter corpus, etc. (ibid.). Besides the list of related word pairs (similar words and human associations), a list of unrelated words has been built. The aim was to build a classifier with the highest average precision in classifying positive and negative samples. In the task of association prediction, positive samples referred to human associations and negatives samples to random words combined to stimuli. The focus of prediction in both tasks were single-word nouns.

The best results for associations from Sociation.org were obtained by combining decision trees based on n-grams, morphological features and neural network models implemented in Word2Vec.[19] Regarding association norms from Russian Associative Lexicon, the best results of association prediction were scored by Word2Vec trained on a combination of ruWac (a subset of Russian National Corpus), Russian Wikipedia and texts from Lib.ru (ibid.). The

[18]More information about the Dialogue conference can be found at http://www.dialog-21.ru/ (03.11.2015).

[19]Word2Vec is an implementation of neural network models (continuous bag-of-words and skip-gram architectures) for computing vector representations of words. It is freely available at https://code.google.com/p/word2vec/ (03.12.2015).

participants used the same models for predicting both similar words as well as human associations. The systems performed better at modelling human associations than synonyms, hypernyms or hyponyms (ibid.), an outcome which may suggest that distributional models as well as recent neural models implemented in Word2Vec are more adequate for predicting words related by means other than strict similarity. However, since the participants of RUSSE task did not analyse the nature of semantic relations between associative pairs before prediction, it remained unknown which kind of semantic relations were covered by the results of the models. In RUSSE there was no attempt to predict human associations only with direct co-occurrence statistics. Contrary to the present investigation, in RUSSE human associations were used as an alternative resource for evaluation of common approaches to semantic similarity, not as an objective of investigation. Accordingly, the reasons for successful or failed predictions were not further discussed.

3 Material

In the following sections I will explain the choice of Russian associations I have chosen to analyse in this study, as well as the choice of corpus that will be used for extracting co-occurrence data.

3.1 Russian association norms

In order to investigate the overlap of co-occurrences with Russian associations, I used the association norms from Slavic Association Dictionary (Ufimceva *et al.* 2004). The choice of Slavic Association Dictionary is motivated by the fact that it contains association norms not only for Russian, but also for Belarusian, Bulgarian and Ukrainian, which are useful for exploring the extent to which the semantic relations within associations are language-independent. The Slavic Association Dictionary contains responses that have been given to 112 stimuli, each represented in four languages (with al small exception for Ukrainian, for which there are 113 stimuli). This makes the chosen lexicon a very valuable resource for exploring the consistency of the association rules discussed in the previous section, which explained that the association norms are interlingual (Raible 1981) or, contrarily, that their language-independence is arguable (Anstatt 2008).

Beside small exceptions,[20] the choice of stimuli is identical for each language. The survey for collecting Slavic association norms was conducted between 1998 and 1999. The participants were students of various disciplines (mathematics, physics, biology, philosophy, science of law, medicine and others), aged 18 to 25 years from the Universities in Moscow, Kursk, Minsk, Sofia and Nizhyn.

The authors of Slavic Association Dictionary gave no precise information about the exact number of participants, probably because the number of participants varied depending on language and on the given stimulus.[21] In retrospect, thanks to the number of responses

[20] Russian has no entries for the stimuli дрэва (Belarusian: wood), чичо (Ukrainian: uncle/older male friend), каханне (Belarusian: love), кола and дума (Bulgarian: car; concept);
Belarusian has no entries for the stimuli брат (Russian: brother), вечность (Russian: eternity), чичо (Ukrainian: uncle/older male friend), кола and дума (Bulgarian: car; concept);
Bulgarian has no entries for the stimuli дрэва (Belarusian: wood), чичо (Ukrainian: uncle/older male friend), женщина (Russian: wife) and каханне (Belarusian: love);
Ukrainian has no entries for the stimuli брат (Russian: brother), вечность (Russian: eternity), кола and дума (Bulgarian: car; concept). (D.B.)

[21] See Ufimceva *et al.* 2004:11: не менее 500 человек (мужчины и женщины примерно в равном количестве), *more than 500, (approximately of both genders equally)* (D.B.).

reported for each stimulus word, I calculated the average number of participants for each language - there were in average 588,7 participants for Russian, 641,5 for Belarusian, 580,2 for Bulgarian and 480,2 for Ukrainian.

The participants of the associative experiment were given a form containing stimuli and were asked to write down the first thing that comes into their mind after having read a stimulus. In order to avoid the priming effect as much as possible, each participant has been given a list with a randomised order of the stimuli. The time for filling the form was limited to 10 minutes. For each language the associations (responses) for each stimulus were collected and counted. An example of an entry (stimulus and all its responses) is given in Figure 3.1. This stimulus in the given entry is the uppercase written word родной (dear, close[22]) and the responses to that stimulus are all the following words and expressions written in order of their frequency, starting from the most frequent to the less frequent association(s).

Figure 3.1: Example entry in Slavic Association Dictionary

РОДНОЙ: язык **195**; *дом* **79**; *человек* **67**; *город* **24**; близкий **23**; *брат* **20**; чужой **19**; край **18**; любимый, отец **13**; мой **12**; свой **6**; папа **5**; *семья*, сын **4**; дорогой, *друг*, милый, *ребенок* **3**; злой, иностранный, неродной, нет, родимый, сестра, страна, теплый, ты мой **2**; а как же, аул, берег, близкий человек, близко душе, близость, быть, *веселый*, двоюродный, деревья, *добро*, дорогой человек, *дядя*, *есть*, козел, конвой, кровь, *лес*, *любовь*, мама, мир, *муж*, национальный, Наш дом Россия, не всегда рад, нежность, необъятный, одна кровь, он же; отец, мать...; очаг, плохой, приемыш, *родина*, родители, родник, *родной*, родственник, родственники, Россия, сват; свой, необходимый, понятный; силуэт, собственность, стиль, счастливый, сынишка, *умный*, университет, хороший; человек, край; человек, муж; что-то свое **1**; ***592+81+7+53+17***

As shown in Figure 3.1, in addition to the list of all the responses and their respective frequencies, an information about the total count of all responses for each stimulus (592) has been given, as well as the total count of all different responses for each stimulus (81), the total count of participants who offered no response to a given stimulus (7), the total count of all responses that appeared only once (53) and the total count of all the responses which are also stimuli in the given form (17).

The stimuli in Slavic Association Dictionary, are language- , genre- (and even time-) independent and it may be assumed that they correspond to the concepts which participants

[22]This and all the other English translations of associations and co-occurrences in this work are given by D.B.

acquired early in life (family members, colours, dimensional adjectives, etc.).[23] As a consequence, these stimuli words are expected to occur with a high frequency in both spoken and written language, which is a premise for the investigation of their contextual distributions.

Nevertheless, like several other collections of association norms (Kent & Rosanoff, South Florida Word Association Norms), Slavic Association Dictionary does not dispose of an equal distribution of stimuli according to their parts-of-speech. No conjunctions, prepositions and particles were given as cue words in the list of stimuli. Instead, one can observe a clear predominance of nouns (63,4%), followed by adjectives and verbs, which are represented in a much smaller proportion (17,9%; 14,3%), and finally by adverbs and numerals, which appear to be very rare (3,6%; 0.9%).[24] For this reason, the analysis of the distributions of word categories is expected to produce most representative results for noun stimuli.

3.2 Corpus material

The corpus I used for estimating the co-occurrence counts consists of the texts collected in the division Современная русская проза (Russian modern literature) of Lib.Ru.[25] I collected the data by choosing one random book for each of 219 Russian writers listed under the division Современная русская проза (modern Russian literature). After the preprocessing, the corpus consisted of 15'553'111 tokens (units divided by white space) and 556'864 types (different tokens). The corpus provides enough text material in order to get reliable estimates of word co-occurrences: a corpus larger than 10 million words is considered to be a decent basis for co-occurrence statistics (cfr. Rapp & Wettler 1991). Texts from Lib.ru have already been used for association prediction within the RUSSE task, where they were combined with Wikipedia texts and a sublet from Russian National Corpus. In this work, I did not mix Lib.ru with other corpora since it can be assumed that literature texts are a sufficient basis for co-occurrence extraction, given that they cover a wide range of topics and contain both direct and indirect speech, which reflects the use of spoken and written language and every-day communication, as well as other registers. Wikipedia texts are, on the contrary, written with a strict scientific style and even though their advantage might have been their large corpus size, they do not represent an ideal basis for extracting co-occurrences that could match human associations

[23] The complete list of Russian stimuli and their English translations is given Table 8.1 in Attachment.
[24] Out of 112 stimuli, there are 71 nouns (for Ukrainian: 72), 20 adjectives, 16 verbs, 4 adverbs and 1 numeral.
[25] http://lib.ru/PROZA/ (16.10.2015).

from Slavic Association Dictionary. An undoubtedly legitimate corpus choice for the aim of this thesis would have been the Russian National Corpus, containing approximately 350 million words, but it can be accessed exclusively through an online interface and the raw corpus data is not available.

4 Analysis of Russian associations

In the following sections I will present the process of retrieving, annotating and analysing the content of Russian associations from Slavic Association Lexicon. The main focus will be the analysis of parts-of-speech and semantic relations within associative pairs. The characteristics of Russian associations will also be compared to the characteristics of Belarusian, Bulgarian and Ukrainian. The questions I aim to answer in the analysis of Russian associations are 1) in which proportions do paradigmatic, respectively, syntagmatic relations occur? 2) in which proportion are similarity, contrast, syntagmatic and semantic contiguity present? 3) to what extent can the distributions of semantic relations within associations be described as language-independent? and 4) which language-related factors determine the distribution of semantic relations within Slavic associative pairs?

4.1 Data gathering

The content of Slavic Association Dictionary is available in printed version as well as on the project's website.[26] In order to gather the content of Slavic Association Dictionary I crawled the project's website by using the Python library BeautifulSoup.[27] I retrieved all the stimuli, their respective responses, the frequency counts of each response as well as other counts related to an associative pairs (number of different answers, number of all answers, etc.). Since in the continuation of this work the associative pairs would be confronted with one-word co-occurrences, I retrieved only one-word responses. Hence, when referring to ten most frequent (or top ten) associations, I refer to the ten most frequent *one-word* associations. The impact of this procedure is discussed in Section 4.2.5.

In association norms taken from Slavic Association Dictionary, it is common to encounter multiple responses having the same frequency. When responses having the same frequency overpassed the limit of ten first associations to consider, a random choice from the set of responses having the same frequency has been taken. For instance, the Russian stimulus дядя (uncle) has the order of responses illustrated in Table 4.1. There were three candidates for the tenth response (мужчина *man*, Федя *Fedja* and чужой *foreign*), out of which the response мужчина was the random choice and was considered further in the analysis.

[26] The book as well as its content in HTML version is available at http://it-claim.ru/Projects/ASIS/SAS/ (15.10.2015).

[27] Free available at http://www.crummy.com/software/BeautifulSoup/ (15.10.2015).

Table 4.1: Ten most common responses for the Russian stimulus **дядя** (- stands for not considered responses)

Considered responses	Rank	Response	Frequency (out of 592)
1	1	тетя *aunt*	134
2	2	Степа *Stepa*	56
3	3	Вася *Vasja*	37
4	4	родственник *relative*	35
5	5	Ваня *Vanja*	31
6	6	родной *dear, close*	16
7	7	мой *my*	15
8	8	Петя *Petja*	12
9	8	друг *friend*	9
10	9	мужчина *man*	8
-	9	Федя *Fedja*	8
-	9	чужой *foreign*	8

In order to facilitate the access to each individual stimulus and its most frequent responses I stored all the available information for ten most frequent responses to each stimulus in a PostgreSQL database.[28] The unique identifiers for all the stimuli from Slavic Association Dictionary were taken from the website. They consist of the language identifier (1 for Russian, 2 for Belarusian, 3 for Bulgarian and 4 for Russian), the initial letter of the stimulus and the number of the stimulus. The unique identifier for responses was produced by adding the rank of the response to the unique identifier of the respective stimulus.

Although the web crawling has proven to be an efficient method for gathering all the relevant data, it has also presented several issues caused by an inconsistent storage of information on the website: the unique identifiers did not always match with the corresponding links,[29] there were several misspelled words[30] or other information mixed with responses,[31] mixed

[28] http://www.postgresql.org/ (15.10.2015).

[29] For instance for 102 instead of 1o2 for Russian stimulus обман (deceit), 2d1 instead of 2r1 for Belarusian stimulus работа (work), etc.

[30] For instance, задасць instead of радасць (joy) in Belarusian data.

[31] For instance, #1088;однина presumably instead of роднина (relative) as response to the Bulgarian stimulus чичо (uncle).

entries among languages,[32] missing of the data for several stimuli,[33] or encoding issues.[34] In order to make the linguistic analysis as accurate as possible, I have manually corrected the mentioned errors and added the missing information before adding the data into the database.

4.2 Procedure

In order to investigate the distribution of semantic relations in associations norms from Slavic Association Dictionary, I firstly considered the presence of paradigmatic and syntagmatic relations and, secondly, the presence of semantic relations presented by Raible (1981), i.e. similarity, contrast, and syntagmatic and semantic contiguity. In order to be able to establish automatically whether there is a paradigmatic or a syntagmatic relation between associative pairs, a part-of-speech tagging was performed, since (in a psychological and computational view of this structuralist dichotomy) the homogeneity of parts-of-speech indicates the paradigmatic, and the heterogeneity the syntagmatic relation.

I analysed the ten most frequent responses for each stimulus, which makes a total of 1120 Russian associative pairs and 3370 Belarusian, Bulgarian and Ukrainian annotated associative pairs. The latter three were used for an inter-language comparison of associations.[35] The choice of ten most frequent responses seemed to be a reasonable choice for investigating the distribution of semantic relations within associative pairs in depth, since the ten most frequent responses accounted for almost the half of all Russian responses.[36]

After the annotation, the database was enriched with meta-information concerning parts-of-speech and semantic relations within associative pairs. When all information was collected, I used SQL queries to measure the distribution of semantic relations for all languages. In order to investigate whether the primary response substantially differs from the other nine responses, I measured the distribution of semantic relations once for the primary responses separately and once for all the ten most frequent responses. Next, I compared the distribution of semantic relations for the Russian language with the distribution of semantic relations within

[32] Russian responses for the stimulus мужчина (man) are erroneously shown as Belarusian responses as well (19.11.2015).

[33] For instance, data for Russian stimuli вечност (eternity) and брат (brother) were available only in the book version of Slavic Association Dictionary (19.11.2015).

[34] Russian letters c, p and x were encoded as Latin letters. The encoding issues originating from the website of Slavic Association Dictionary were resolved with help of Mirjam Zumstein (Email communication from 09.10.2015).

[35] As mentioned before, for Ukrainian there was one stimulus more, making a total of 1130 associative pairs.

[36] Ten most frequent responses accounted in average for a total of 47.2% Russian responses and 43.3% for all four languages.

associations from other languages. The results were related to previous studies of associations. To conclude the analysis of Russian association norms, the possibility of their prediction with methods of distributional semantic has been discussed.

4.2.1 Part-of-speech tagging

For the part-of-speech annotation of Russian and Bulgarian association norms I used the statistical part-of-speech tagger TreeTagger.[37] Since the parameter files for Belarusian and Ukrainian are not (yet) incorporated into TreeTagger, I annotated the parts-of-speech of associative norms for these two languages manually. In order to be able to compare the distribution of word classes among four different languages, I used the international tagset proposed by Petrov and colleagues (2011). As shown in Table 4.2, the international tagset consists of 12 tags. Since the TreeTagger for Russian and Bulgarian does not assign the categories from the international tagset, but Russian[38] and Bulgarian[39] specifications, I converted the output from TreeTagger into the categories from the international tagset.[40] Since the analysis of parts-of-speech is fundamental for the present investigation, I have manually corrected the errors.[41] In case of ambiguities (for instance, the Russian stimulus больной *sick/sick person* might be interpreted as adjective as well as a noun) I assigned the part-of-speech which, to my judgement, was the one that most of the participants would associate first (e.g. in case of the stimulus больной (ill) it would be an adjective instead of a noun).

Considering that I intended to work with a lemmatised Russian corpus in the prediction part of this work, I also used Tree Tagger to lemmatise Russian responses, so that they could be compared to the lemmatised textual co-occurrences.

[37]The TreeTagger is a tool for annotating part-of-speech and lemma (main word form) information. It was developed by Helmut Schmid at the Institute for Computational Linguistics in Stuttgart. It is freely available at http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/ (24.10.2015).

[38]The Russian tagset (Morphosyntactic Specifications for Russian) was developed by Sharoff and colleagues within the project MULTEXT-East. A table with the explanation of all the tags is available at http://corpus.leeds.ac.uk/mocky/ru-table.tab (24.10.2015).

[39]Bulgarian tagset (BTB-TR03: BulTreeBank Morphosyntactic Tagset) was developed by Kiril Simov and colleagues (2004) within the BulTreeBank project.

[40]The words have been fed into the TreeTagger separately, since no context is provided. By doing so, TreeTagger had assigned the tags that were the most frequent for the respective word (e.g. if in the training corpus the same word was tagged as both noun and verb, but more frequently as noun than a verb, TreeTagger assigns the noun tag). The performance of the TreeTagger for both Russian and Bulgarian was satisfying: once the tags were converted into the international tagset, in 112 tagged stimuli I used as a test set, I have found only 2 erroneously tagged words for Russian and 3 for Bulgarian.

[41]The proportion of corrected international tags was 2% for Russian (22 out of 1120) and 3% (37 out of 1120) for Bulgarian.

Table 4.2: International tagset proposed by Petrov *et al.* (2011)

Tag	Explanation
VERB	verbs (all tenses and modes)
NOUN	nouns (common and proper)
PRON	pronouns
ADJ	adjectives
ADV	adverbes
ADP	adpositions (prepositions and postpositions)
CONJ	conjunctions
DET	determiners
NUM	numerals
PRT	particles or other function words
X	other: foreign words, typos, abbreviations
.	punctuation

4.2.2 Annotation of semantic relations

The annotation of contrast, similarity, semantic contiguity and syntagmatic contiguity was done manually. Since Raible's (1981) and Anstatt's (2008) criteria for assigning semantic relations between associative pairs do not always coincide and since the two researchers do not discuss the way they have used to disambiguate problematic cases, I have defined several criteria for annotating the type of semantic relation.

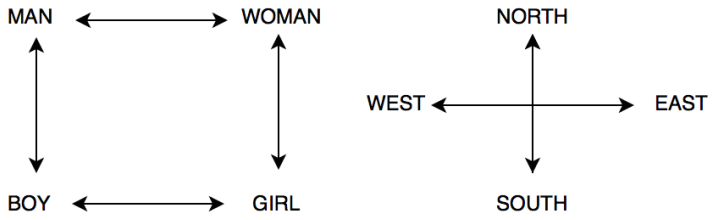
Contrast As a criterion for defining the contrast I have used the antonymy in its wide sense,[42] as both Raible and Anstatt did. The types of antonyms I considered were scalar antonyms (long-short), complementary antonyms (man-woman),[43] and converses (buy-sell). Another antonymy type I considered was a "borderline collection of antonyms" type (Jones

[42]"In its broadest sense, antonymy covers a wide range of word pairs expressed by different part-of-speech categories as long as these words express the opposite of each other." Lobanova 2012:10)

[43]As Jones highlights, it may be argued that gender opposites are not opposites at all, since in many European languages the "opposite" pair *boy* and *girl*, effectively remains the same gender marked word: Italian has *ragazzo* and *ragazza*, Spanish has muchacho and muchacha, Portuguese has *menino* and *menina* (2002:18). Although the meaning remains the same, it may appear intuitive to classify *boy* and *girl* as opposites, but *ragazzo* and *ragazza* as similar. In the observed Slavic association norms I did not encounter gender opposites which differentiated one from another only in morphology.

2002:18) called multiple incompatibility, including pairs such as *winter-summer*, *north-south*.[44] In the case of multiple incompatibility, I differentiated between "orthogonal" and "antipodal" oppositions (Lyons 1977:286), illustrated in Figure 4.1.

Figure 4.1: Orthogonal (left) and antipodal opposition (right) (Lyons 1977), source: Jones (2002:20)



As shown in Figure 4.1, the orthogonal opposition describes the antonymy between the words such as *man, woman, girl* and *boy*, where *man* is as antonym of *boy* and *woman* but not of *girl*, and *woman* is antonym of *man* and *girl*, but not of *boy*. In the antipodal opposition words contrast only in one direction, like *north* to *south* and *west* to *east*. In assigning the contrast relation, I considered all the relations in orthogonal and antipodal oppositions except the vertical relations as *woman-girl*. I did not consider as contrast the relation between co-hyponyms having no direct oppositions, such as days of a week or months. According to the mentioned criteria, Russian associative pairs I have marked as opposites are, for instance, белый-черный (white-black), бабушка-дедушка (grandmother-grandfather), друг-враг (friend-enemy), день-ночь (day-night), ненавидеть-любить (to hate-to love), терять-находить (to lose-to find), огонь-вода (fire-water) and земля-небо (earth-sky).

As "there is nothing to prohibit polyantonymy" (Jones 2002:17), I considered pairs счастье-горе (happiness-sorrow) as well счастье-несчастье (happiness-unhappiness) as antonym pairs. In order to measure the presence of contrast relation adequately, I annotated for each stimulus whether it (theoretically) had an opposite word or not. By doing so, I differentiated among three levels of opposition strength the respective stimulus could display, ranging from *high level* of antonymy, referring to stimuli which might have any sorts of antonyms among adjectives, *medium level* of antonymy referring to stimuli of other parts-of-speech (complementary

[44] According to Jones (2002:18) in some respects, multiple incompatibility is most similar to non-gradable antonymy. The non-gradable pair *female* and *male*, for example, belongs to a two-member system, such that X can never be simultaneously more than one member: *solid*, *liquid* and *gas*, by comparison, belong to a three-member system, such that X can never be simultaneously more than one member; similarly, *clubs*, *diamonds*, *hearts* and *spades* belong to a four-member system, such that X can never be simultaneously more than one member.

antonyms, converse, reciprocal antonyms) and *low level* of antonymy, referring to stimuli which might have an antonym only in the widest sense of antonymy, for instance душа-сърце (soul-heart) or душа-тяло (soul-body). I made this differentiation because firstly, the definition of antonymy is widely discussed in literature, and secondly, because I wanted to investigate whether a contrast association depended on the antonym type.[45]

Similarity For what concerns similarity, I considered the *is-a* relations (synonymy, near-synonymy, hyponymy and hyperonymy), but also incompatible co-hyponyms such as красный-синий (red-blue).[46] Examples of similarity are Бог-Христос (God-Christ), дурак-идиот (fool-idiot), вода-жидкостъ (water-liquid), хлеб-еда (bread-food), красный-синий (red-blue) and радость-веселье (joy-cheeriness).

Syntagmatic contiguity I used the Raible's term syntagmatic contiguity to refer to the pairs which had an agreement (congruence) as тёмные-волосы (dark-hair) and начало-дня (beginning-day[GEN.SG.]); case government, as читать-книгу (to read-book[ACC.SG.]); and verb complements, as попросить-позвонить (to ask[somebody]-to call). At this point, it must be remembered that there is a difference between the notion of syntagmatic relation referring to the associative pairs having heterogeneous parts-of-speech and the notion of syntagmatic contiguity, which refers only to the associative pairs having heterogeneous parts-of-speech, which are also syntactically dependent of one another.

Semantic contiguity Following the example of Raible (1981) and Anstatt (2008), I marked all the associative pairs which were not connected by the relations of similarity, contrast and syntactic contiguity as falling into the category of semantic contiguity. Some examples are лицо-зеркало (face-mirror), любовь-секс (love-sex), работа-деньги (work-money), муж-любовник (husband-lover).[47]

According to the criteria explained above, the semantic relations between the Russian stimulus бабушка (grandmother), as well as for all the other stimuli from Slavic Association

[45]While some researches prefer to use the term antonyms to refer to scalar adjectives only (Lyons 1977; Cruse 1986), others claim that the term antonyms should be used more generally (Jones 2002).

[46]The compatible co-hyponyms, like гост-приятел (Bulgarian: guest-friend) which may both be interpreted as co-hyponyms, since the common hyperonym is *person*, were not considered as similar.

[47]Although this criterion may seem simplistic, the semantic contiguity can be considered as a generic term for describing many different kind of semantic relations which may hold between two words which are related by any kind of common context.

Dictionary and their most frequent responses were assigned as in Table 4.3.

Table 4.3: Semantic relations between the stimulus **бабушка** and its ten most frequent responses

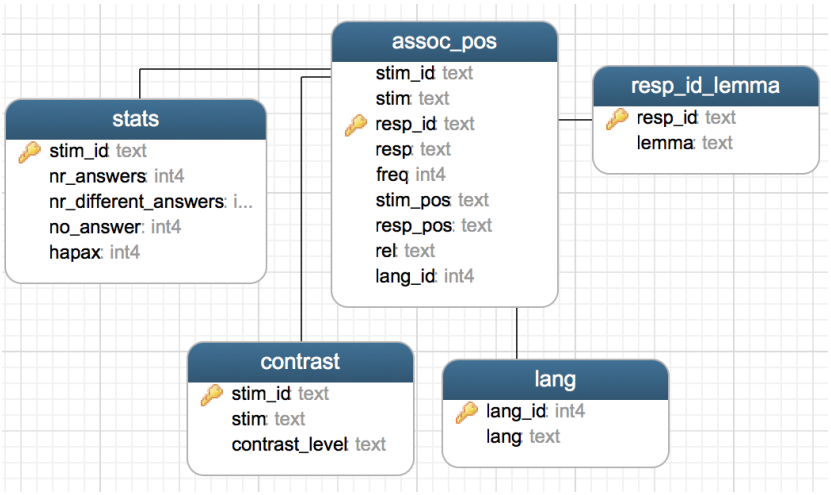
Frequency (out of 592)	Response	Semantic relation
128	дедушка *grandfather*	contrast
54	старая *old*	synt. contiguity
35	добрая *good*	synt. contiguity
29	старушка *old woman*	similarity
16	моя *my*	synt. contiguity
11	любимая *dear*	synt. contiguity
11	пирожки *cookies*	sem. contiguity
9	доброта *goodness*	sem. contiguity
9	родная *dear, close*	synt. contiguity
9	старость *maturity*	sem. contiguity

Defining semantic relations to words without having a context is not a trivial task. Firstly, for what adjective stimuli are concerned, it is not possible to determine whether a participant in the associative experiment had interpreted them as being adjective or nouns, since in Russian adjectives may be used as substantives: for example, when the stimulus черный (black) has been presented, a participant may have interpreted it as referring to the colour as well as as referring to a black person. In each case, the relation between the stimulus черный (black) response and the response негр (Negro) would have been different: in case of the associative pair черный-негр, where черный is interpreted as an adjective, the semantic relation to assign would be syntagmatic contiguity, whereas in case черный has been interpreted as an adjective in substantive use, the relation to assign would be similarity. For the cases in which both alternatives were possible, I assigned the semantic relation which in my best judgement was most likely to be the one the participants had in mind as they made the link between stimulus in response (in this case, черный was probably interpreted in a substantive use of adjective, hence, the semantic relation I assigned was similarity). Another challenge was the annotation of semantic relations between stimuli and deictic elements, such as those in associative pairs красивый-я (beautiful-I), молодой-я (jung-I), умный-я (intelligent-I), чалавек-я (Belarusian: man-I). It may be assumed that participants in those cases referred to their own selves. In those cases, I marked the semantic relations as syntagmatic contiguity, since the two words may be combined in a syntagma. However, in the case of чалавек-я (Belarusian: man-I), for instance,

it is also possible to classify the underlying relation as similarity, since the words *I* and *man* can be interpreted as standing in *is-a* relation to one another. A similar ambiguity arises when an *is-a* relation referring to a metaphor appears, such as in глупый-пень (stupid-stump) and глупый-осел (stupid-donkey), where both пень and осел are personifications of human beings. Other examples where the metaphor is encountered are the associative pairs are жыццё-шлях (life-path) and смърт-край (death-end).[48] I did not consider the metaphorical *is-a* relation as similarity, but as either syntactic contiguity or as semantic contiguity, depending on the syntactic compatibility of the words within the respective associative pair.

Once having annotated the associative pairs according to their parts-of-speech and semantic relation, I enriched the database with the corresponding information. After this step, the database had the structure illustrated in Figure 4.2.[49]

Figure 4.2: Graph representing the structure of the database



The associative pairs as well as their parts-of-speech tags and semantic relations were stored into table *assoc_pos* , which was connected to 1) table *stats*, containing all the counts gathered from the website, 2) table *contrast*, which contained the information about the antonymy level of stimuli, 3) Table *resp_id_lemma*, which contained lemmatised responses and 4) Table *lang*, which stored the unique identifiers which have been given for each language.

[48] In this context, it is impossible not to recall Lakoff's work on conceptual metaphors, stating that LIFE IS A JOURNEY, the death, on the other hand, being conceptualized as the end of a journey (1993:224).

[49] Figure 4.2 has been generated by the database administration tool Navicat, available at http://www.navicat.com (19.11.2015).

4.2.3 Calculations

I calculated the relative frequency distributions of parts-of-speech and semantic relations, considering the primary response as well as all ten most frequent responses. For instance, for calculating the distribution of noun responses for noun stimuli, I divided the number of the observed noun responses with the total number of noun stimuli. In order to explore the assumed prevalent role of contrast associations (Raible 1981), I calculated the frequency distributions of semantic relations for the associative pairs which may have a contrast. In addition, I calculated the proportions of *high, medium* and *low* types of antonymy levels in order to observe whether particular antonym types elicit more opposite responses than others. Since psychological studies have shown that different parts-of-speech elicit different types of responses (cfr. Section 2.3), I calculated 1) the proportion of responses which preserve the part-of-speech of the stimulus and 2) the proportion of the semantic relations for each considered part-of-speech (nouns, adjectives and verbs) separately. To access the database from a Python script I used PostgreSQL database adapter Psycopg.[50]

4.2.4 Expectations

According to the previous study of Anstatt (2008), arguing that syntagmatic contiguity appears in a greater proportion than contrast in Slavic languages, it was to expect that the majority of primary responses would be syntagmatic. Accordingly, the part-of-speech of the response is expected to be different than the part-of-speech of the stimulus. The contiguity axis (syntactic and semantic contiguity) is expected to be more pronounced than the similarity axis (contrast and similarity). A similar pattern was expected to be found in all other analysed Slavic languages.

4.2.5 Limitations of the procedure

Disregarding multi-word responses may have slightly influenced the results to the disadvantage of syntagmatic relations, since multi-word responses are often syntagmatic, as бог-на небе (God-in the sky), надеяться-и верить (to hope-and to believe), разговор-ни о чем (conversation-about nothing). In order to estimate the impact of the skipped responses, I have counted the appearance of multi-word responses in the ten most frequent Russian responses.

[50] Psycopg is the most popular PostgreSQL adapter for the Python programming language. It is freely available at http://initd.org/psycopg/ (26.10.2015).

Out of 1120 responses, 52 were multi-word responses, which means that the differences in the distributions of syntagmatic vs. paradigmatic responses were altered up to 4.6%. The effect of ignoring the multi-word responses is, however, certainly not dramatic, since they never occurred as primary response and since their absence might have been compensated by other responses standing in syntagmatic relation with the stimulus.

Given that the criteria for assigning the type of semantic relations are not universal and that I was the only annotator for this experiment, it is important to mention that the disambiguation decisions may vary in an eventual repetition of the experiment.

4.3 Results

Almost two thirds (65.18%) of Russian primary responses from Slavic Association Dictionary preserve the part-of-speech of the stimulus, whereas in 34.82% of the cases the parts-of-speech within most common associative pairs differ from one another, as shown in Table 4.4. When the ten most frequent associations are considered, the distribution of homogeneous associative pairs is almost equal to the distribution of heterogeneous associative pairs (50.45%, respectively, 49.55%).

Table 4.4: Distribution of heterogeneous/homogeneous associative pairs

	Primary response (total 112)	Ten most frequent responses (total 1120)
Heterogeneous (stimulus PoS $\neq$ response PoS)	34.82% (39)	49.55% (555)
Homogeneous (stimulus PoS = response PoS)	65.18% (73)	50.45% (565)

The interaction of homogeneity of associations with the parts-of-speech is shown in Table 4.5. When primary responses are considered, noun stimuli mostly elicit noun responses (73.24%), verb stimuli, likewise, mostly elicit verb responses (62.5%), whereas adjective stimuli combine with responses of other parts-of-speech in 65% of the cases. When ten most frequent responses are considered, noun stimuli still elicit mostly noun responses (61.27%), adjective stimuli still elicit mostly non-adjective responses (77.50%), but verb stimuli mostly combine with non-verb responses (56.88%).

One may observe the difference between responses to noun, adjective and verb stimuli more in detail in Table 4.6. Primary responses to noun stimuli are mostly nouns (73.24%),

Table 4.5: Distribution of heterogeneous/homogeneous associative pairs for nouns, adjective and verbs

	Most frequent associative pair		Ten most frequent associative pairs	
	Homogeneous	Heterogeneous	Homogeneous	Heterogeneous
NOUN (total 71)	73.24% (52)	26.76% (19)	61.27% (435)	38.73% (275)
ADJ (total 20)	35.00% (7)	65.00% (13)	22.50% (45)	77.50% (155)
VERB (total 16)	62.50% (10)	37.50% (6)	43.13% (69)	56.88% (91)

followed by adjectives 18.31% and 5.63% verbs, whereas responses of other parts-of-speech to noun stimuli are clearly underrepresented. Adjective stimuli elicit noun responses in 65% of the cases and adjective responses in 35% of the cases. Other than verb responses, verb stimuli elicit nouns and adverbs in 18.75% of the cases each.

Table 4.6: Distribution of parts-of-speech for noun, adjective and verb stimuli for primary responses

PoS	NOUN	ADJ	VERB	ADV	NUM	PRON
NOUN (total 71)	73.24% (52)	18.31% (13)	5.63% (4)	0.00%	1.41% (1)	1.41% (1)
ADJ (total 20)	65.00% (13)	35.00% (7)	0.00% (0)	0.00% (0)	0.00% (0)	0.00% (0)
VERB (total 16)	18.75% (3)	0.00% (0)	62.50% (10)	18.75% (3)	0.00% (0)	0.00% (0)

When ten most frequent responses are considered (Table 4.7), one can still observe that noun and verb stimuli mostly elicit responses of same part-of-speech, whereas the majority of verb stimuli elicits responses of different parts-of-speech. Nouns stimuli show a more diversified spectrum of parts-of-speech within responses than adjectives and verbs do, since they also combine with adverbs, numerals, pronouns and particles. What adjectives are concerned, when the response is not a noun, it is almost certainly an adjective: 76% of the most frequent ten responses to adjective stimuli are nouns and 22.50% are adjectives. Responses to verb stimuli are mostly verbs (43.14%), followed by nouns (33.13%) and adverbs (15.63%). Numerals, pronouns and particles are very rarely represented, whereas prepositions and conjunctions are never observed within the ten most frequent associations.

According to the results of manual annotation of contrast, similarity, syntagmatic and semantic contiguity, one can observe that syntagmatic contiguity is overall the most frequent semantic relation in Russian associative pairs (37.50%), directly followed by contrast (35.71%), semantic contiguity (19.64%) and similarity (7.14%), as shown in Table 4.8. However, when only stimuli having a contrast are considered, the dominance of syntagmatic contiguity over

Table 4.7: Distribution of parts-of-speech for noun, adjective and verb stimuli for ten most frequent responses

PoS	NOUN	ADJ	VERB	ADV	NUM	PRON	PRT
NOUN	61.27% (435)	27.32% (194)	4.93% (35)	2.68% (19)	0.70% (5)	2.96% (21)	0.14% (1)
ADJ	76.00% (152)	22.50% (45)	0.00% (0)	0.00% (0)	0.00% (0)	1.50% (3)	0.00% (0)
VERB	33.13% (53)	0.63% (1)	43.13% (69)	15.63% (25)	3.13% (5)	4.38% (7)	0.00% (0)

contrast is inverted, the latter being the most represented semantic relation (47.62%), as shown in Table 4.9. If the stimulus is a word that might have an opposite, however, the response is observed to consist of the opposite in almost half of the cases, especially when responses to noun stimuli are considered.

Table 4.8: Distribution of semantic relations between stimuli and primary responses (total 112)

contrast	similarity	sem. contiguity	synt. contiguity
35.71% (40)	7.14% (8)	19.64% (22)	37.50% (42)

Table 4.9: Distribution of semantic relations between stimuli and primary responses if stimulus has a contrast (total 84)

contrast	similarity	sem. contiguity	synt. contiguity
47.62% (40)	3.57% (3)	11.90% (10)	36.90% (31)

The most represented contrast type is the one displaying a *medium* level of antonymy (antonym nouns and verbs), for which for 28 out of 43 possible cases (65.11%) contrast was the primary response. *High* contrast level most common associative pairs elicited were encountered in 10 out of 24 possible cases, whereas low *level* of contrast was observed in 2 out of 17 possible cases.

When considering semantic relations between ten most frequent responses to each stimulus, syntagmatic contiguity accounts for almost a half of all semantic relations (47.41%), followed by semantic contiguity (34.38%), similarity (10%) and contrast (8.04%), as shown in Table 4.10. Phonetic similarity has proven to be a very rare semantic relation in associative pairs (0.18%). The only associative pairs, in which phonetic similarity seemed to be the underlying semantic relation, are дочь-ночь (daughter-nigth) and гость-кость (guest-bone).

As shown in Table 4.11 noun stimuli mostly elicit opposite responses (36.62%), whereas

Table 4.10: Distribution of semantic relations holding between Russian stimuli and ten most frequent responses (total 1120)

contrast	similarity	sem. contiguity	synt. contiguity	phon. similarity
8.04% (90)	10.00% (112)	34.38% (385)	47.41% (531)	0.18% (2)

adjectives and verbs mostly elicit responses which are syntagmatically contiguous with the stimulus (65% and 37.50%, respectively). Responses to nouns and verbs adopt a different distributive behaviour than the responses to adjective stimuli, which display less variation in the distribution of semantic relations. In fact, if one re-examines the Table 4.6, one can observe that each time an adjective stimulus has elicited an adjective response, it has always been a contrast.

Table 4.11: Distribution of semantic relations for primary responses to stimuli of different parts-of-speech

PoS	contrast	similarity	sem. contiguity	synt. contiguity
NOUN (total 71)	36.62% (26)	9.86% (17)	23.94% (7)	29.58% (21)
ADJ (total 20)	35.00% (7)	0.00% (0)	0.00% (0)	65.00% (13)
VERB (total 16)	25.00% (4)	6.25% (1)	31.25% (5)	37.50% (6)

When ten most frequent responses are considered, semantic contiguity is the predominant relation for noun stimuli (41.41%), whereas syntagmatic contiguity dominates for both adjective and verb stimuli (70.50%, respectively, 48.75%), as shown in Table 4.12.

Table 4.12: Distribution of semantic relations within ten most frequent associative pairs (total 1120)

PoS	contrast	similarity	sem. contiguity	synt. contiguity
NOUN (total 710)	7.32% (52)	11.13% (79)	41.41% (249)	39.86% (283)
ADJ (total 200)	10.50% (21)	8.00% (16)	11.00% (22)	70.50% (141)
VERB (total 160)	7.50% (52)	8.75% (14)	35.00% (56)	48.75% (78)

4.4 Comparison with other Slavic languages

As it is the case for Russian associations, an overall preference towards homogeneous responses can be observed for Belarusian and Bulgarian, where the part-of-speech of the stimulus equals to part-of-speech of primary responses in 57.14%, respectively, 77.68% cases. In contrast,

Ukrainian associations are mostly of heterogeneous nature: homogeneous associative pairs account for only 25.66% of all associative pairs. When ten most frequent responses are considered, the prevalence of homogeneous associative pairs diminishes for Russian, Belarusian and Bulgarian, and increases for Ukrainian, as shown in Table 4.13.[51]

Table 4.13: Distribution of homogeneous associative pairs for Russian (ru), Belarusian (be), Bulgarian (bg) and Ukrainian (uk)

Language	Primary response (total 112)	Ten most frequent responses (total 1120)
ru	65.18% (81)	50.45% (565)
be	57.14% (72)	52.95% (593)
bg	77.68% (95)	60.80% (681)
uk	25.66% (36)	37.17% (420)

Regarding noun stimuli there is a common tendency for noun responses for Russian (73.24%), Belarusian (77.46%) and Bulgarian (87.32%), whereas Ukrainian noun stimuli elicit mostly adjective responses (64.38%), as shown in Table 4.14.[52] The particularity of Ukrainian can also be observed what adjective stimuli are concerned, where 100% of the responses (20 out of 20) are always nouns. In Russian and Belarusian, adjective stimuli elicit mostly noun responses as well (65.00%; 90.00%), but Bulgarian adjectives mostly elicit other adjectives (55.00%). When stimulus is a verb, Russian and Bulgarian show a tendency towards verbs responses (62.50%; 62.50%), whereas Belarusian and Ukrainian verb stimuli mostly attract noun responses (50%; 43.75%). As it is the case in Russian, noun stimuli attract the most varied span of responses for other languages as well, whereas adjective stimuli elicit either other adjectives or nouns. Pronoun and numeral responses are observed only as responses to noun stimuli, whereas conjunctions, particles and prepositions are not represented as primary responses for any language. Although several common tendencies have been observed, there is no precise common rule to describe the behaviour concerning the distribution of parts-of-speech within primary responses which would account for all four languages. However, when ten most frequent responses are considered (cfr. Table 4.15), one may observe that regardless of the part-of-speech of the stim-

[51] In order to ascertain that the collection of association norms for Ukrainian was made under the exact same conditions as for the other languages, I contacted Galina Čerkasova, the co-creator of Slavic Association Dictionary. Čerkasova assured that the task was exact the same for all languages (Email communication dating 24.09.2015), which let us assume that there were no difficulties in data gathering which would compromise the quality of the collection. It would be interesting to investigate whether this peculiarity of Ukrainian language would persist in other associative experiments as well.

[52] For space reasons, the observed values for 4.14 as well as for 4.15 are reported in Tables 8.2 and 8.3 in Attachment.

ulus, the responses appear to be mostly nouns, the only exception being Russian, for which verb responses tend to prevail over noun responses (43.13% vs. 33.13%).

Table 4.14: Distribution of part-of-speech tags for noun, adjective and verb stimuli in Russian, Bularusian, Bulgarian and Ukrainian primary responses

		NOUN	ADJ	VERB	ADV	NUM	PRON
NOUN	ru	73.24%	18.31%	5.63%	0.00%	1.41%	1.41%
	be	77.46%	22.54%	0.00%	0.00%	0.00%	0.00%
	bg	87.32%	9.86%	1.41%	0.00%	1.41%	0.00%
	uk	32.88%	64.38%	0.00%	1.37%	0.00%	0.00%
ADJ	ru	65.00%	35.00%	0.00%	0.00%	0.00%	0.00%
	be	90.00%	10.00%	0.00%	0.00%	0.00%	0.00%
	bg	45.00%	55.00%	0.00%	0.00%	0.00%	0.00%
	uk	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
VERB	ru	18.75%	0.00%	62.50%	18.75%	0.00%	0.00%
	be	50.00%	0.00%	31.25%	18.75%	0.00%	0.00%
	bg	31.25%	0.00%	62.50%	6.25%	0.00%	0.00%
	uk	43.75%	0.00%	31.25%	25.00%	0.00%	0.00%

Table 4.15: Distribution of part-of-speech tags for noun, adjective and verb stimuli in Russian, Bularusian, Bulgarian and Ukrainian ten most frequent responses

		NOUN	ADJ	VERB	ADV	NUM	PRON
NOUN	ru	61.27%	27.32%	4.93%	2.68%	0.70%	2.96%
	be	68.17%	25.77%	2.54%	1.27%	0.00%	2.11%
	bg	77.04%	16.06%	2.96%	2.39%	0.56%	0.99%
	uk	47.95%	42.74%	3.97%	1.92%	0.41%	1.64%
ADJ	ru	76.00%	22.50%	0.00%	0.00%	0.00%	1.50%
	be	73.50%	25.00%	0.00%	0.00%	0.00%	1.00%
	bg	57.50%	39.00%	0.00%	0.00%	0.50%	1.00%
	uk	86.00%	13.50%	0.00%	0.00%	0.00%	0.50%
VERB	ru	33.13%	0.63%	43.13%	15.63%	3.13%	4.38%
	be	53.75%	1.88%	31.88%	9.38%	2.50%	0.63%
	bg	47.50%	1.25%	28.75%	10.63%	5.00%	6.88%
	uk	53.13%	0.00%	23.13%	20.00%	1.88%	1.88%

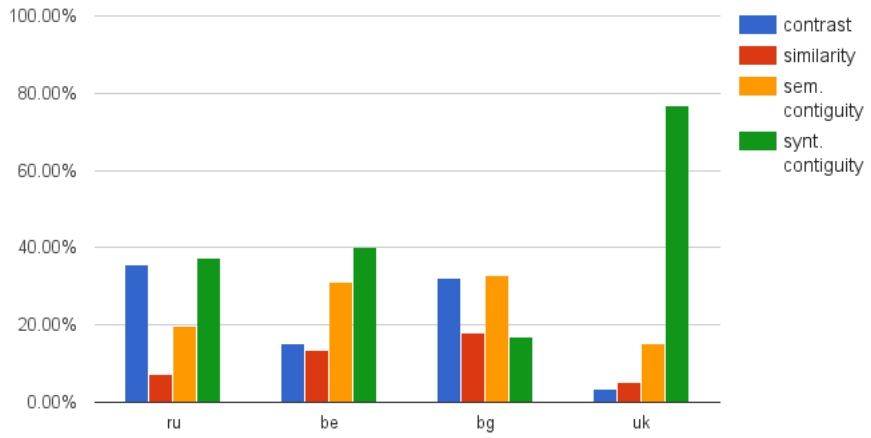
Regarding the distribution of semantic relations, one may observe that they are distributed for each language differently, as shown in Table 4.16 and illustrated in Figure 4.3: syntagmatic contiguity dominates in Russian (37.50%), Belarusian (40.18%) and especially in Ukrainian (76.79%) primary responses, whereas the dominant relation in Bulgarian associative pairs is semantic contiguity (33.04%). The contrast relation is the second most commonly represented

for Russian and Bulgarian associative pairs (35.71%; 32.14%), but it is the third and, respectively, the rarest relation for Belarusian (15.18%) and Ukrainian (3.57%). The similarity is overall the least represented relation within primary responses, whereas the phonological similarity is, as expected, never observed.

Table 4.16: Distribution of semantic relations in primary responses for Russian, Belarusian, Bulgarian and Ukrainian (total 112)

	contrast	similarity	sem. contiguity	synt. contiguity
ru	35.71% (40)	7.14% (8)	19.64% (22)	37.50% (42)
be	15.18% (17)	13.39% (15)	31.25% (35)	40.18% (45)
bg	32.14% (36)	17.86% (20)	33.04% (27)	16.96% (19)
uk	3.57% (4)	5.36% (6)	15.18% (17)	76.79% (86)

Figure 4.3: Distribution of semantic relations in primary responses for Russian, Belarusian, Bulgarian and Ukrainian



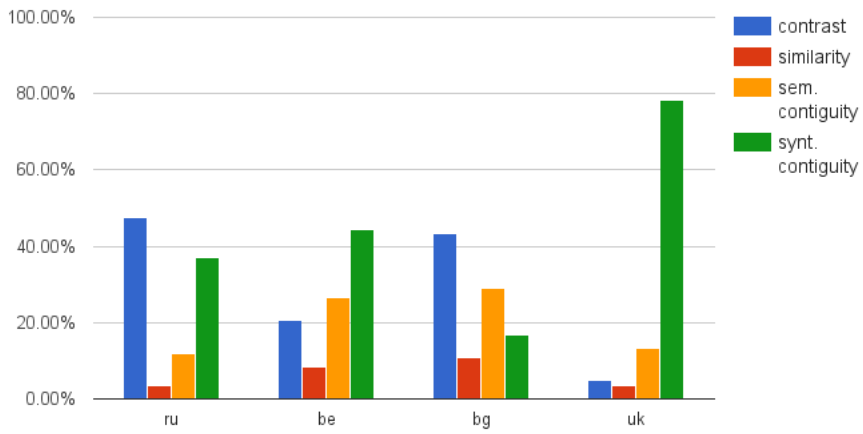
If one considers only the stimuli which have a contrast, the distribution of the semantic relations looks slightly different: the contrast takes over the predominance of syntagmatic contiguity for Russian and of semantic contiguity for Bulgarian, as shown in Table 4.17 and illustrated in Figure 4.4. Taking into consideration only stimuli which have a contrast slightly affects the distribution of Belarusian and Ukrainian as well, but since in these languages the contrast association (as well as similarity) is underrepresented, the increase in contrast relation

does not have an impact on the predominance of syntagmatic contiguity.

Table 4.17: Distribution of semantic relations for primary response if response has a contrast (total responses: 84 for Russian, 83 for other languages)

	contrast	similarity	sem. contiguity	synt. contiguity
ru	47.62% (40)	3.57% (3)	11.90% (10)	36.90% (31)
be	20.48% (17)	8.43% (7)	26.51% (22)	44.58% (37)
bg	43.37% (36)	10.84% (9)	28.92% (24)	16.87% (14)
uk	4.82% (4)	3.61% (3)	13.25% (11)	78.31% (65)

Figure 4.4: Distribution of semantic relations for primary response if response has a contrast (total responses: 84 for Russian, 83 for other languages)



When the ten most common associative pairs are considered, syntagmatic contiguity dominates for Russian and Ukrainian (47.41%; 59.82%) whereas semantic contiguity dominates for Belarusian and Bulgarian (42.86%; 49.46%). Differences in distribution of types of antonyms that elicit the most contrast responses can also be observed: while for Russian and Belarusian stimuli the *medium* level of antonymy (antonym nouns and verbs) is the most represented type of contrast (28/43; 13/43) and the only one represented for Ukrainian (4/43), the *high* level of antonymy (antonym adjectives) is the most represented type of contrast for Bulgarian (15/24), as shown in Table 4.18.

For stimuli having a contrast, the contrast association is approximately the third ranked

Table 4.18: Differences in predominant contrast relations in Russian, Belarusian, Bulgarian and Ukrainian primary responses

Language	Non-binary antonyms (*low* level of antonymy)	Antonyms within nouns or verbs (*medium* level of antonymy)	Antonyms within adjectives (*high* level of antonymy)
ru	11.76% (2/17)	65.12% (28/43)	41.67% (10/24)
be	5.88% (1/17)	30.23% (13/43)	12.50% (3/24)
bg	16.67% (3/17)	42.86% (18/43)	62.50% (15/24)
uk	0.00% (0/17)	9.30% (4/43)	0.00% (0/24)

response for Russian and Bulgarian, the fourth most frequent for Belarusian and the fifth most frequent for Ukrainian.[53] Similarly as it has been observed for Russian, in Belarusian, Bulgarian and Ukrainian the primary response never accounts for more than 50% of overall responses, and only in few cases for more than a quarter of all responses.[54]

All summed up, establishing prevalences of semantic relations that are valid for all four languages appears to be a challenging task. However, one may observe that all the languages show certain common tendencies: the primary responses are generally either in relation of syntagmatic contiguity with the stimulus, or are opposite in meaning, whereas responses similar to the stimulus word are the least represented for all languages. The syntagmatic contiguity is the most frequent primary response when all stimuli are taken into consideration. When only the stimuli having a contrast are considered, both syntagmatic responses as well as contrast responses are represented, but their dominance is language-dependent: while Ukrainian and Bulgarian show a tendency towards syntagmatic responses, Russian and Bulgarian are similar in distribution because they both prefer contrast relations. As shown in Table 4.19, one may observe that only Bulgarian follows the hierarchy of associations proposed by Raible (43.37% contrast, 28.92% semantic contiguity, 16.87% syntagmatic contiguity and 10.84% similarity) whereas none of the languages follows the hierarchy observed by Anstatt (2008), although for Belarusian and Ukrainian the syntagmatic contiguity is the most frequent semantic relation in associative pairs.

[53]The average number of contrast association is 2.9 for Russian, 3.95 for Belarusian, 3.1 for Bulgarian and 4.9 for Ukrainian.

[54]For Belarusian and Bulgarian there are 7 associative pairs for which more than 25% of the participants agreed, for Ukrainian 6.

Table 4.19: Hierarchy of associative responses according to Raible (1981), Anstatt (2008) and observed hierarchy (obs.) of Russian, Belarusian, Bulgarian and Ukrainian responses for stimuli which have a contrast

	Raible's hierarchy	Anstatt's hierarchy	obs. hierarchy Russian resp.	obs. hierarchy Belarusian resp.	obs. hierarchy Bulgarian resp.	obs. hierarchy Ukrainian resp.
1	contrast	synt. contiguity	contrast	synt. contiguity	contrast	synt. contiguity
2	sem. contiguity	contrast	synt. contiguity	sem. contiguity	sem. contiguity	sem. contiguity
3	synt. contiguity/	sem. contiguity/	sem. contiguity	contrast	synt. contiguity	contrast
4	similarity	similarity	similarity	similarity	similarity	similarity

4.5 Discussion

According to the findings of this analysis, if *syntagmatic* is understood as heterogeneous (referring to the presence of different parts-of-speech between stimulus and response) and *paradigmatic* as homogeneous, the distribution of paradigmatic and syntagmatic relations within associative pairs varies with the part-of-speech of the stimulus: what Russian most frequent associative pairs are concerned, noun and verb stimuli tend to elicit paradigmatic responses (73.24%; 62.50%) whereas adjectives tend to elicit syntagmatic responses (65%). While responses to noun and verb stimuli vary in their distribution of parts-of-speech of responses as well as of the distribution of semantic relations (such as contrast, similarity and syntagmatic/semantic contiguity), adjective stimuli elicit almost exclusively noun or adjectives responses and are connected to the stimulus by either contrast of syntagmatic contiguity.

The observations regarding the importance of parts-of-speech within the associative pairs imply that the prevalence of a semantics relation in particular association norms depends on the content of the list of stimuli: if stimuli are nouns, the prevalence of paradigmatic relations within primary associative pairs is to be expected, if stimuli are adjectives, the prevalence of syntagmatic relations between associative pairs can be assumed, but if stimuli are verbs, the prevalence of one or the other might be expected. The overall distribution of paradigmatic/syntagmatic dichotomy, as well as the distribution of semantic relations within associations is, therefore, relative to the proportion of parts-of-speech within the stimuli. Hence, in contrary to what has been hypothesised by Schulte im Walde & Meligner (2008), an investigation based only on verb stimuli cannot be generalised to all the other parts-of-speech with the assumption that they all display a similar behaviour.

The more responses are taken into consideration, the more the observations regarding parts-of-speech in the present analysis seem to support the findings of Ervin (1961), Deese (1692), Cramer (1968) and Palermo (1971), who claimed that noun stimuli tend to attract

more paradigmatic responses than other parts-of-speech as well, as those of De Deyne & Storms (2008), who reported that noun responses are the most frequent even if stimuli are adjectives or verbs.

According to what has been hypothesised by Raible (1981), the rule: "if the word has an opposite, the association will be an opposite word" still holds in case of Russian associations: it has been observed that, if the stimulus has a contrast, in the majority of the cases, the response to that stimulus has been a contrast (47.62%). If the response is a contrast, then it is in a paradigmatic relation with the stimulus. Therefore, the proportion of paradigmatic relations depends not only on the part-of-speech of the stimulus, but also on the number of words in the list of stimuli which might have a contrast: the more words with a possible contrast word are present, the more contrast relations, and, respectively, the more paradigmatic relations are expected to be found within associative pairs. Contrary to what has been hypothesised by Anstatt (2008), syntagmatic contiguity is only the second most represented semantic relation between Russian stimuli which might have a contrast and their responses (36.90%), but it is the most represented one when responses to all stimuli are considered (37.50%).

The distribution of semantic relations, depends, however, not only on the choice of words within the list of stimuli, but it is also language-specific; as it has been observed in comparison with Belarusian, Bulgarian and Ukrainian associations, although there are common tendencies, the distribution of semantic relations within associative pairs differs for all considered languages. If a stimulus has a contrast, the response is prevalently a contrast what Russian and Bulgarian language are concerned, but the syntagmatic contiguity prevails for Belarusian and Ukrainian associations. The only constant seems to be the fact that similarity is poorly represented in all investigated languages.

Whether the distribution of semantic relations within Slavic associations indeed differs from their distribution in Germanic languages, as it has been claimed by Anstatt (2008), cannot be determined on the basis of the results of this study, since the parallel associative experiment has not been performed in the same time span and with the same list of stimuli in Germanic-speaking countries. However, it can be hypothesised that hypothetical Germanic associative pairs would display more paradigmatic relations than the Slavic ones, and thus for a very simple reason: Slavic languages have a different morphology characteristics which are likely to influence the choice of responses. While adjectives in Germanic languages might be

presented in an unmarked form in the list of stimuli, there is no unmarked form for adjectives in Slavic languages: for instance, the adjective *sauber* (clean) in German is not marked and thus the participant of the associative experiment can interpret it as being an adverb or as being an adjective. In both cases, since contrast association is observed to be very frequent, the response might be expected to be a contrast. However, in Slavic languages, for instance in Russian, for the adjective *sauber* (clean) there is no correspondent unmarked form: the translation of *sauber* in Russian nominative can be either чистый[ADJ.M], чистая[ADJ.F] or чистое[ADJ.N], in Slavic Association Dictionary the masculine form чистый has been chosen. The German equivalent would be *sauber*[ADJ.] or *sauberer*[ADJ.M.]. It is legitimate to hypothesise that syntagmatic responses to adjective stimuli, which have been observed in 65% of Russian cases, are due to the completion of the noun chunk which has been initiated with a masculine adjective, similarly as it would have possibly been for German associative pairs analysed in Anstatt (2008) if the German adjective stimulus would have been presented in the masculine form. Anstatt (2008:27) has already mentioned the morphology as possible reason for the highly represented syntagmatic responses in Slavic associations. However, she did not consider the fact that stimuli in Slavic languages were in masculine form as a crucial reason for the overall prevalence of syntagmatic associations, with the explanation that even in the cases where the stimuli were not marked (adverbs), Russian responses tended to give more syntagmatic than contrast responses, and that even verbs in unmarked, infinite form, tended to elicit syntagmatic responses (2008:29-30). As shown in the analysis of Russian associations in this work, this is not the case either for Russian or for Bulgarian associations in Slavic Association Dictionary.[55]

Although it might be said that Raible (1981) overestimated the role of contrast by stating that "whenever possible, almost without exceptions"[56] a stimulus having an opposite word will elicit its opposite, it may as well be said that Anstatt (2008) underestimated the presence of contrast associations in Slavic associations. Moreover, one may also argue that she overstated the possible unity of Slavic languages what semantic relations within associative pairs are concerned by establishing the prevalence of syntagmatic contiguity merely for Russian and

[55]In Russian and Bulgarian associations from Slavic Association Dictionary, three out of four adverb stimuli were contrast: for Russian быстро-медленно (fast-slow), плохо-хорошо (bad-good) and хорошо-плохо (good-bad); for Bulgarian бързо-бавно (fast-slow), лошо-добро (bad-good), добре-зле (good-bad). In both languages, verb elicit mostly paradigmatic primary responses.

[56]"Assoziationen sind, wo immer dies möglich ist, fast ausnahmslos Assoziationen zum Entgegengesetzten hin" (1981:12). D.B.

Polish languages.

Now that the distribution of semantic relations within associative pairs in Slavic Association Dictionary has been determined, it has been made possible to compare it with the distribution of semantic relations in text co-occurrences as well as to investigate the semantic features of the overlap between associations and co-occurrences further in this work.

5 Analysis of the overlap

In the following sections I will present the procedure I used to analyse the overlap between co-occurrences and human associations, and the results of the analysis. The focus of the analysis was, firstly, the lexical overlap between co-occurrences and associations, secondly, the overlap in semantic relations between co-occurrences and associations, and thirdly, the proportion of semantic relations within the co-occurrences overlapping with human associations. I am going to refer to co-occurrences with highest log likelihood score as primary co-occurrences or *predicted* primary responses.

5.1 Procedure

Before extracting co-occurrences, I cleaned the corpus by eliminating punctuation as well as meta-information originating from the Lib.ru website and by normalizing word accentuation. After the cleaning phase, the corpus contained only lowercased Cyrillic words. In order to extract co-occurrences from the corpus I implemented the log likelihood algorithm in a Python script.[57] I calculated co-occurrences of the corpus by filling the contingency table described in Section 2.2.1. For each Russian stimulus listed in Slavic Association Dictionary I extracted ten co-occurrences with the highest log likelihood score. I experimented with the following parameters:

1. context size: To investigate the effect of the context size on the overlap statistics I measured the overlapping statistics for the co-occurrences within context windows of ±2, ±5, ±10 and ±20 words. Testing different context windows is motivated by the fact that different context window may extract different types of co-occurrences.
2. lemmatised/unlemmatised corpus: for a morphologically rich language such as Russian, the lemmatisation, i.e. the process of normalizing all words in the corpus to their main word forms, may be of great importance for the performance of association measures. For lemmatising the corpus I used TreeTagger; and
3. part-of-speech selection: since conjunctions and adpositions (prepositions and postpositions) had never appeared among the analysed Russian associations, I tested the differ-

[57] Although mutual information is the oldest and most widely used association measure (Paperno et al. 2014), according to Lezius (1999) and Evert & Krenn (2001), its precision is significantly lower than that of log likelihood. Moreover, in the evaluation of Evert & Kenn (2001), log likelihood was found it to be one of the best association measures.

ence in overlap between co-occurrences and associations with and without presence of conjunctions and adpositions.[58]

For each parameter combination, I measured the overlap between co-occurrences and human associations by considering the ten most frequent associations, primary associations as well as all associations. Co-occurrences obtained from the unlemmatised corpus were compared with unlemmatised human associations, whereas those obtained from the lemmatised corpus were compared with the lemmatised human associations.

I analysed the part-of-speech distribution as well as the distribution of semantic relations. As it was the case in part-of-speech tagging of human associations, all co-occurrences have been provided with an international part-of-speech, which has been assigned by adapting the tags given by TreeTagger. The semantic relations were principally assigned according to the same criteria as human associations. However, there were two main differences which I had to consider; the first consists in the fact that the connection between stimulus and common responses is almost always traceable, whereas the connection between co-occurrences calculated by log likelihood is not, as in the examples гость-стол (guest-table), человек-она (man-she).[59] In these cases, I marked the type of relation as *unknown*. The second difference relates to the co-occurrences originating from the lemmatised corpus, such as such as дочь-младший (daughter-younger[ADJ.M.]) or дочь-старший (daughter-older[ADJ.M.]), where the gender of the adjective does not match the gender of the noun, since the agreement between the two has been neutralized by lemmatisation. Despite the absence of agreement, I marked the relation between those co-occurrences as syntagmatic contiguity. An example of the annotated co-occurrences of the stimulus дверь (door) is given in Table 5.1.[60]

The results of the models were fed into the database and queried by means of PostgreSQL database adapter Psychopg.

[58]The list of Russian conjunctions was taken from https://en.wiktionary.org/wiki/Category:Russian_conjunctions (16.10.2015), the list of adpositions from http://www.study-languages-online.com/grammar/tables/prepositions-by-case (16.10.2015). Although the list of adpositions was diversified and contained 76 items, particular forms of adpositions, in which for phonetic reasons the suffix (о) is often added, such as обо *about* and изо (out of), were not present in the list.

[59]The examples originate from the model trained on a lemmatised corpus with a context window of ±20 words.

[60]The examples originate from the model trained on a lemmatised corpus with a context window of ±20 words.

Table 5.1: Co-occurrences with the highest log likelihood score (LLR), part-of-speech (PoS) and semantic relation for the stimulus word **дверь**

Co-occurrence	LLR	PoS	Semantic relation
открыть *to open*	3156.4874	VERB	syn. contiguity
коридор *corridor*	2441.262	NOUN	sem. contiguity
комната *room*	2162.4177	NOUN	sem. contiguity
войти *enter*	1788.7353	VERB	sem. contiguity
постучать *to knock*	1465.0715	VERB	sem. contiguity
запереть *to lock*	1442.4808	VERB	synt. contiguity
порог *threshold*	1337.142	NOUN	sem. contiguity
распахнуть *to unbar*	1285.7172	VERB	synt. contiguity
замок *lock*	1090.9576	NOUN	sem. contiguity
закрыть *to close*	1083.4238	VERB	synt. contiguity

5.2 Expectations

Since Russian is a morphologically rich language, the lemmatised corpus was expected to produce more accurate co-occurrences and, hence, more overlaps with the human associations than the unlemmatised corpus. Regarding the context window, I assumed that the context window of ±5 would catch the majority of co-occurrences which correspond to associations, since it is possibly large enough to capture relations of semantic contiguity, which are not necessarily in direct proximity, but not so large to neutralize the effect of strict proximity, responsible for capturing relations of syntagmatic contiguity, which have proven to be frequent within Russian associations.[61] Accordingly, it was expected that the smaller the window, the more syntagmatic contiguity relations, the larger the window, the less syntagmatic contiguity and more semantic contiguity. Similarity was not expected to be represented within co-occurrences, since similar words tend to share common contexts rather than occur together. Contrast relations were expected to occur more frequently that similarity, but less frequently than syntagmatic contiguity. Given that co-occurrence extraction benefits from the syntagmatic use of context, the co-occurrences were not expected to preserve the part-of-speech of the stimulus, but to be prevalently heterogenous.

[61]Similar was claimed by Church and Hanks, who argued that the context window of 5 words was large enough to "show some of the constraints between verbs and arguments, but not so large that it would wash out constraints that make use of strict adjacency" (1990:24).

5.3 Results

The highest overlap between co-occurrences and human associations was achieved by the model trained on the lemmatised corpus with the context size of ±20 words, in which conjunctions and adpositions were not considered. The model with this parameter combination extracted primary co-occurrences which corresponded to primary responses in 22.32% of the cases (cfr. Table 5.3). In comparison, the average participant in the Russian associative experiment from Slavic Association Dictionary produced 17,3 primary responses to 112 stimuli (meaning that, in average, 15.5% responses of an average participant coincided with the most frequently given response). The overlap between the ten most frequent co-occurrences and the ten most frequent responses amounted to 20.98% (cfr. Table 5.2). The overlap between primary co-occurrences of the stimulus word and any human response to the stimulus word accounted to 68% for the context window of ±20 words, and 70% for the context window of ±10 words. For the overview of all the overlaps for the model trained on the lemmatised corpus with the context window of ±20 words see Table 8.4 in Attachment.[62]

Table 5.2: Overlap between the ten most frequent associations to a stimulus and ten most frequent co-occurrences of the stimulus (absolute counts out of 1120 are written in parenthesis)

Relative overlap	Context window ±2	Context window ±5	Context window ±10	Context window ±20
unlemmatised without PoS selection	15.09% (169)	16.25% (182)	16.34% (183)	15.36% (172)
lemmatised without PoS selection	14.64% (164)	18.57% (208)	19.29% (216)	19.82% (222)
unlemmatised with PoS selection	15.45% (173)	16.52% (185)	16.61% (186)	15.71% (176)
lemmatised without PoS selection	16.88% (189)	19.73% (221)	20.18% (226)	20.98% (235)

As it can be observed in Tables 5.2 and 5.3, the overlap of co-occurrences and human associations depends on the context window: the larger the context window, the better the overlap. For each context window the proportion of overlap depended on the lemmatisation, the only exception being the context of ±2 words in the measurement of the overlap between ten most frequent associations and ten co-occurrences. The larger the context window, the

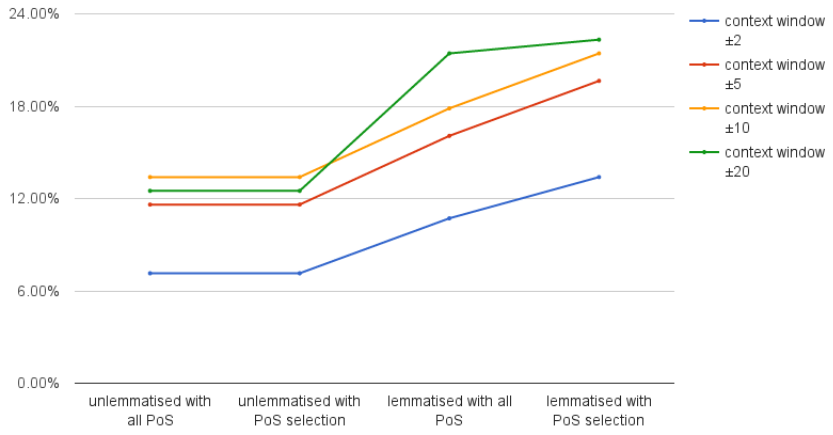
[62]The proportion of overlap for the model trained on the lemmatised corpus with the context window of ±2 words is 60% (67 out of 112), for the model trained on the lemmatised corpus with the context window of ±5 words 67% (75 out of 112).

Table 5.3: Overlap of primary responses with primary co-occurrences (absolute values out of 112 are written in parenthesis)

Relative overlap	Context window ±2	Context window ±5	Context window ±10	Context window ±20
unlemmatised without PoS selection	7.14% (8)	11.60% (13)	13.39% (15)	12.5% (14)
lemmatised without PoS selection	10.71% (12)	16.07% (18)	17.86% (20)	21.43% (24)
unlemmatised with PoS selection	7.14% (8)	11.60% (13)	13.39% (15)	12.5% (14)
lemmatised with PoS selection	13.39% (15)	19.64% (22)	21.43% (24)	22.32% (25)

greater the impact of the lemmatisation on the overlap proportion, especially what the overlap of primary co-occurrences with primary responses is concerned, where the overlap is almost half as high when the co-occurrences are extracted from the lemmatised corpus (cfr. Table 5.3: context window of ±20 unlemmatised without PoS selection: 12.5%; context window of ±20 lemmatised without PoS selection: 21.43%). The impact of the lemmatisation on the overlap between primary co-occurrences and primary responses for each context window is depicted in Figure 5.1. As it can be observed, the part-of-speech selection did not show to have such a great impact on the overlap as the lemmatisation did; moreover, what the overlap between primary co-occurrence and primary responses is concerned, it did not make any difference. This observation indicates that co-occurrences found by the log likelihood ratio do not contain a high amount of function words such as prepositions and conjunctions. Since the overlap for lemmatised models with the part-of-speech selection was slightly higher than without the selection, when considering the overlap statistics further in this work, I will refer to the results without conjunctions and adpositions, unless explicitly indicated otherwise.

Figure 5.1: Overlap between primary co-occurrences and primary responses in function of lemmatisation and PoS selection



In comparison to co-occurrences extracted by models trained on the lemmatised corpus, co-occurrences found by the models trained on the unlemmatised corpus contained multiple different word forms having the same meaning. For example, consider the co-occurrences of the stimulus дверь (door) in Table 5.4. The co-occurrences extracted from the unlemmatised corpus contained two different forms of the verb открыть (to open), which are открыл (opened[PFV.PST.SG.M.IRREFL])[63] and открыла (opened[PFV.PST.SG.F.IRREFL]) within the top ten co-occurrences.[64] The co-occurrences from the lemmatised model, instead, contained only one main word form, which allowed other co-occurrence candidates to land within the top ten co-occurrences and compete for the overlap with human responses.[65]

Although the overlap of human associations with co-occurrences in the example of дверь is not high, the co-occurrences extracted by both models may intuitively be thought of as words which one may associate to the stimulus *door*, such as *corridor, open, close, knock, lock,*

[63]The standard I use for labelling linguistic information are The Leipzig Glossing Rules (Lehmann 1982).

[64]The extraction of multiple word forms having the same meaning are also frequent for noun stimuli, for example for the stimulus друг (friend) within ten most frequent co-occurrences one also observe другу (friend [DAT.]) and другом (friend [INS.]).

[65]For the sake of simplicity, I did not enrich the translations with the morphological information.

Table 5.4: Ten most frequent responses to the word **дверь** (door) in comparison with their ten most frequent co-occurrences extracted from the results of the models trained on the unlemmatised and lemmatised corpus with the context window of ±20 words

Ten most frequent associations	Frequency (out of 592)	Ten most frequent co-occurrences (unlemmatised corpus)	LLR		Ten most frequent co-occurrences (lemmatised corpus)	LLR	Presence in ten most frequent associations
открыта *opened*	58	открыл *opened*	1724.46	-	открыть *to open*	3156.48	-
окно *window*	50	открылась *opened*	1444.62	-	коридор *corridor*	2441.26	-
закрыта *closed*	29	в *in*	1104.87	-	комната *room*	2162.41	-
ручка *handle*	29	распахнулась *unbarred*	1065.31	-	войти *enter*	1788.73	-
замок *lock*	22	хлопнула *slammed*	1021.21	-	постучать *to knock*	1465.07	-
вход *entrance*	20	открыла *opened*	947.31	-	запереть *to lock*	1442.48	-
деревянная *wooden*	18	отворилась *opened*	866.17	-	порог *swell*	1337.14	-
дубовая *oak-*	17	распахнул *unbarred*	678.96	-	распахнуть *to unbar*	1285.71	-
выход *exit*	16	постучали *knocked*	615.88	-	замок *lock*	1090.95	x
дом *home*	16	собой *itself*	563.08	-	закрыть *to close*	1083.42	x

etc.[66] One of the stimuli words for which the highest overlap of top ten responses and top ten co-occurrences was observed is the stimulus родной (dear/close), for which six out of ten co-occurrences matched with any of top ten human responses, as shown in Table 5.5.

Other stimuli words which had at least 50% or more overlap in at least one model are пить (to drink), зеленый (green) and брат (brother). According to the results reported in Table 5.6, the number of overlaps within ten most frequent co-occurrences and ten most frequent responses for these stimuli varied from four to six. In comparison, the average number of overlap within top ten co-occurrences and top ten responses for each stimulus was 2.03 for

[66]When comparing the co-occurrences of дверь with the responses from the book entrances of Slavic Association Dictionary one can notice that the verb открыть (to open) is, indeed, the tenth most frequent association to дверь and that it should be considered as an overlap. The unrecognised overlap is due the fact that the number of responses to consider was limited to ten, and when there were numerous responses with the same frequencies competing for the tenth place a random response was chosen as the tenth response to consider in the analysis, as explained in Section 4.1.

Table 5.5: Ten most frequent responses to the word **родной** (dear/close) in comparison with its ten most frequent co-occurrences extracted from the results of the models trained on the unlemmatised and lemmatised corpus with the context window of ±20 words

Responses	Frequency (out of 592)	Co-occurrences	LLR	Presence in ten most frequent associations
язык *language*	195	брат *brother*	332.04	x
дом *home*	79	мать *mother*	257.95	-
человек *man*	67	близкий *close*	213.42	x
город *town*	24	дом *home*	207.97	x
близкий *close*	23	город *town*	203.25	x
брат *brother*	20	отец *father*	192.01	x
чужой *foreign*	19	язык *language*	183.73	x
край *region*	18	сестра *sister*	163.09	-
любимый *dear*	13	земля *country*	145.39	-
отец *father*	13	сын *son*	137.80	-

models trained on the unlemmatised corpus and 2.25 for models trained on the lemmatised corpus.[67]

Table 5.6: Stimuli which always resulted with high overlap between co-occurrences and responses

Stimulus	Unlem ±2	Lem ±2	Unlem ±5	Lem ±5	Unlem ±10	Lem ±10	Unlem ±20	Lem ±20
родной	5	6	5	6	4	6	4	6
пить	4	5	5	5	4	4	4	4
зеленый	5	4	4	4	4	4	4	4
брат	5	5	5	6	5	5	5	4

The stimulus родной, which reached the highest overlap between co-occurrences and responses was also the stimulus with the highest variance in human responses (24.34, in comparison to an average of 9.11).[68] However, for particular stimuli there was no single overlap in any co-occurrences: for instance, no co-occurrences of the stimuli Бог (God) and время (time) corresponded to any of the top ten most common human responses to those stimuli.

[67]For unlemmatised models, the averages were 1.99 for the context window of ±2, 2.09 for the context window of ±5, 2.08 for the context window of ±10 and 1.98 for the context window of ±20. For lemmatised models, the averages were 2.05 for the context window of ±2, 2.34 for the context window of ±5, 2.32 for the context window of ±10 and 2.31 for the context window of ±20.

[68]Moreover, родной is the stimulus to which the participants in the associative experiment responded with the lowest number of different responses (81), which compares to an average of 178.96 different responses for all Russian stimuli.

The variance of responses to бог is interestingly one of the lowest (4.26).[69]

In the results of almost all models, verb stimuli generally had a lower overlap between responses and co-occurrences than stimuli belonging to other parts-of-speech (cfr. Table 5.7). In the results of the model trained on the lemmatised corpus with the context window of ±20, one fifth of all verbs had no single overlap. The number of adjective stimuli displaying no overlap within the top ten co-occurrences amounts only to 5% for the the model trained on the lemmatised corpus with the context window of ±20. Noun stimuli have no single overlap in 9.8% of the cases for the best model.

Table 5.7: Parts-of-speech of stimuli which never had an overlap in responses and co-occurrences

	PoS	Context window ±2	Context window ±5	Context window ±10	Context window ±20
Unlemmatised	NOUN (71)	18.30% (13)	18.30% (13)	19.70% (14)	18.30%% (13)
	ADJ (20)	10.00% (2)	10.00% (2)	10.00% (2)	10.00% (2)
	VERB (16)	50.00% (8)	43.70% (7)	37.50% (6)	43.70% (7)
Lemmatised	NOUN (71)	22.5% (16)	15.49% (11)	12.67% (9)	9.80% (7)
	ADJ (20)	10.00% (2)	10.00% (2)	10.00% (2)	5.00% (1)
	VERB (16)	37.50% (6)	31.25% (5)	37.50% (6)	25.00% (4)

The distributions of parts-of-speech within co-occurrences are shown in Tables 5.8 and 5.9.[70] As in the case of human associations, when top ten co-occurrences are considered (Table 5.8), co-occurrences to noun stimuli are mostly paradigmatic (more than 50% of the responses to noun stimuli are nouns), whereas co-occurrences of adjective stimuli are mostly syntagmatic (only 23.5% responses to adjective stimuli are other adjectives). The same applies to the distribution of part-of-speech tags within primary co-occurrences (Table 5.9): noun stimuli mostly co-occur with other nouns (53.52%), whereas most common co-occurrences of adjective stimuli are nouns as well (50%), which makes their prevalent relations syntagmatic. However, the distribution of parts-of-speech within the co-occurrences of verb stimuli differs from those in human associations in the case when the ten most common co-occurrences are

[69]The lowest variance is given within responses to stimuli человек (man), обещать (promise) and думать (think) (3.48; 4.003; 4.005), which also had a low overlap statistics: at most two overlaps between responses and co-occurrences of the stimulus человек and at most one for the stimuli обещать and думать.

[70]For a visual representation of the Tables regarding the distribution of parts of speech in co-occurrences and associations see Figures 8.1 and 8.2 in Attachment.

considered, as well as in the case when only primary co-occurrences are considered.[71]

Table 5.8: Distribution of parts-of-speech for noun, adjective and verb stimuli for ten most frequent responses vs. ten most frequent co-occurrences extracted with model trained on lemmatised corpus with the context window of ±20

PoS	NOUN	ADJ	VERB	PRON	other
NOUN (10 co-occurrences)	50.84% (361)	13.38% (95)	18.73% (133)	12.53% (89)	4.50% (32)
NOUN (10 responses)	61.26% (435)	27.32% (194)	4.92% (35)	2.95% (21)	3.52% (25)
ADJ (10 co-occurrences)	54.00% (108)	23.5% (47)	3.50% (7)	10.5% (21)	8.50% (16)
ADJ (10 responses)	76.00% (152)	22.50% (45)	0.00% (0)	1.50% (3)	0.00% (0)
VERB (10 co-occurrences)	39.37% (63)	0.62% (1)	23.12% (37)	18.75% (30)	18.12% (29)
VERB (10 responses)	33.12% (53)	0.62% (1)	43.12% (69)	4.37% (7)	18.75% (30)

Table 5.9: Distribution of parts-of-speech for noun, adjective and verb stimuli for primary responses and primary co-occurrences extracted with model trained on lemmatised corpus with the context window of ±20

PoS	NOUN	ADJ	VERB	PRON	other
NOUN (primary co-occurrences)	53.52% (38)	9.86% (7)	18.31% (13)	14.08% (10)	4.23% (3)
NOUN (primary responses)	73.24% (52)	18.31% (13)	5.63% (4)	0.00% (0)	2.82% (2)
ADJ (primary co-occurrences)	50.00% (10)	40.00% (8)	5.00% (1)	5.00% (1)	0.00% (0)
ADJ (primary responses)	65.00% (13)	35.00% (7)	0.00% (0)	0.00% (0)	0.00% (0)
VERB (primary co-occurrences)	50.00% (8)	0.00% (0)	6.25% (1)	37.50% (6)	6.25% (1)
VERB (primary responses)	18.75% (3)	0.00% (0)	62.50% (10)	0.00% (0)	18.75% (3)

The major difference seems to consist in the fact that co-occurrences for all stimuli, and particularly in regard to verb stimuli, often consist of pronouns, which are not as much represented in human associations, especially within primary responses (0 out of 112 primary responses is a pronoun). The high proportion of pronouns within co-occurrences may explain why the verb stimuli have less overlaps than noun and adjective stimuli, as it was already mentioned above. For example, one may observe the co-occurrences of the stimulus думать (to think), which, as mentioned above, never display an overlap with most common human responses (see Table 5.10).

Table 5.10 represents the variety of parts-of-speech which may be encountered in top ten co-occurrences since they occur with the stimulus more frequently than it would be expected by chance (pronouns and particles), however, they are not words which people tend to associate

[71] At this point, I must stress that the output of the part-of-speech tagger was not manually checked, so that the distribution of the parts-of-speech is to be considerate as approximate and not absolute. Moreover, it also must be remembered that the distribution in Tables 5.8 and 5.9 regards only the co-occurrences extracted by the model trained on the lemmatised corpus with the context window of ±20 words, which proved to overlap the best with human data what top ten co-occurrence and top ten responses are concerned.

Table 5.10: Co-occurrences of **думать** extracted from the output of the model trained on the lemmatised model with the context window of ±20 words

Responses	Frequency	Co-occurrences	LLR	Overlap
мыслить *to think*	27	я *I*	3886.22	-
много *a lot*	26	ты *you*[PRON.2.SG]	2093.53	-
мечтать *to daydream*	25	не *not*	1343.33	-
размышлять *to think*	18	так *so*	661.96	-
долго *long*	18	вы *you*[PRON.2.PL]	518.88	-
мысль *thought*	15	это *that*	510.77	-
мысли *thoughts*	13	вот *here*	322.14	-
решать *to solve*	12	нет *no*	308.90	-
хорошо *good*	11	ну *well*	261.89	-
быстро *fast*	10	мысль *thought*	257.32	x

when a content word stimulus such as noun, verb or adjective is presented.

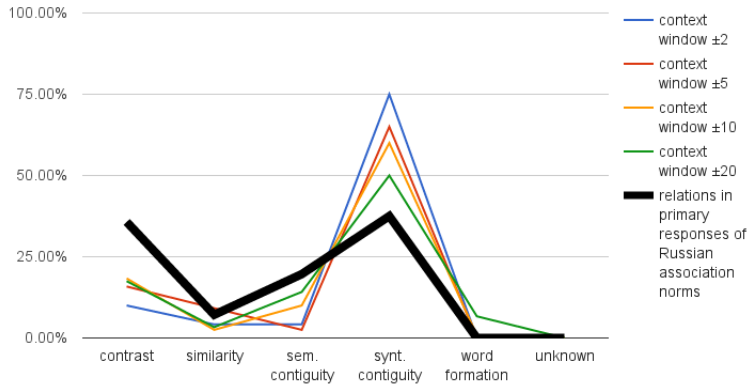
The distribution of semantic relations between stimuli and primary co-occurrences mirrors the distribution of semantic relations between stimuli and primary responses. Just like in the case of human associations, in the majority of co-occurrences syntagmatic contiguity is the most frequent semantic relation, directly followed by contrast relation, whereas semantic contiguity and similarity are the least frequent. As shown in Table 5.11, the wider the context window the more the distribution of semantic relations within co-occurrences and stimuli resembles the distribution of semantic relations between associative pairs. As it can be observed, the model trained with the context window of ±20 words produces not only a higher overlap with human primary responses, but with 50% syntagmatic contiguity, 17.50% contrast, 14.17% semantic contiguity and 3.33% similarity it also has the most similar distribution of semantic relations as human associative responses. The effect of context size on the distribution of semantic relations is depicted in Figure 5.2. Within the co-occurrences extracted from the results of the models trained on context windows of ±10 and ±20 words one can also observe co-occurrences connected by the relation of word formation, for instance, жить-жизнь (to live-life) and любовь-любить (love-to love), which do not appear among human primary responses.

Although the distribution of semantic relations within co-occurrences nears the distribution of semantic relations within human responses, the overlap between the two is based mostly on one semantic relation, namely, on contrast. This means that the most of the co-occurrences which were correctly predicted, i.e. which overlapped with human associations, were in a

Table 5.11: Proportion of semantic relations between stimuli and their primary co-occurrences for context windows of ±2, ±5, ±10 and ±20 words in comparison with the proportion of semantic relations between stimuli and their most frequent Russian responses

	contrast	similarity	sem. contiguity	synt. contiguity	word formation	unknown
context ±2	10.00% (12)	4.17% (5)	4.17% (5)	75.00% (90)	0.00%	0.00%
context ±5	15.83% (19)	9.17% (11)	2.50% (3)	65.00% (78)	0.00%	0.00%
context ±10	18.33% (22)	2.50% (3)	10.00% (12)	60.00% (72)	0.83% (1)	1.67% (2)
context ±20	17.50% (21)	3.33% (4)	14.17% (17)	50.00% (60)	6.67% (8)	1.67% (2)
Russian primary responses	35.71% (40)	7.14% (8)	19.64% (22)	37.50% (42)	0.00%	0.00%

Figure 5.2: Proportion of semantic relations between stimuli and their primary co-occurrences for context windows of ±2, ±5 and ±10 words in comparison with the proportion of semantic relations within most frequent Russian associative pairs



contrast relation with the stimulus word. This observation regards especially the overlap within primary co-occurrences and primary responses, the proportion of which can be observed in Table 5.12. Contrary to the expectations, the overlap between primary co-occurrences and primary responses was very rarely based on syntagmatic contiguity (at most in 9.52% of the cases). The proportion of the overlap regarding similarity amounts to 12.15%. The overlap of primary co-occurrences based on semantic contiguity is the second highest with 18.18% overlap with human primary responses.

An example of the overlapping primary co-occurrences and their semantic relations for all context windows is shown in Table 5.13. As it can be observed, the overlap based on

Table 5.12: Semantic relation in overlap between primary responses and primary co-occurrences for unlemmatised and lemmatised models (the percent values refer to the ratio between semantic relation which was captured in the overlap and the number of the respective semantic relation found in Russian primary co-occurrences)

		contrast	similarity	sem. contiguity	synt. contiguity
Unlemmatised	Context ±2	7.50% (3)	12.50% (1)	4.55% (1)	7.14%(3)
	Context ±5	20.00% (8)	12.50% (1)	4.55% (1)	7.14% (3)
	Context ±10	25.00% (10)	12.50% (1)	4.55% (1)	7.14% (3)
	Context ±20	22.50% (9)	12.50% (1)	4.55% (1)	7.14(3)
Lemmatised	Context ±2	27.50% (11)	0.00% (0)	4.55% (1)	7.14% (3)
	Context ±5	40.00% (16)	0.00% (0)	7.14% (2)	9.52% (4)
	Context ±10	45.00% (18)	0.00% (0)	9.09% (2)	9.52% (4)
	Context ±20	40.00% (16)	12.15% (1)	18.18% (4)	9.52% (4)
Russian primary responses		(40)	(8)	(22)	(42)

contrast relation regards even smaller window sizes, which indicates that contrasting words often occur together. The overlap of primary co-occurrences with primary responses which are based on syntagmatic contiguity is very poorly represented; it concerns only the pairs молодой-человек (young-man), свет-яркий (light-bright), умный-человек (intelligent-man) and справедливость-восторжествовать (justice-to triumph). Similarity accounts for one overlap стыд-позор (shame-dishonour), whereas semantic contiguity accounts for four overlaps: искать-найти (to search-to find), лес-дерево (forest-tree) and палец-рука (finger-hand) and радость-счастье (joy-happiness). The larger the context window, the better the predictions for all semantic relations.

When considering the overlap between the ten most frequent co-occurrences and the ten most frequent associations (cfr. Table 5.14), the extension of the context window as well as the lemmatisation do not guarantee a higher overlap for all semantic relations: the associative pairs based on syntagmatic contiguity are captured the best in the results of the model trained on the unlemmatised corpus with the context window of ±2 words (25.99%) and the least represented within the overlap of the model trained on the lemmatised corpus with the context window of ±20 (19.21%), which proved to have the best performance regarding the representation of all the other semantic relations. Accordingly, when the overlap between top ten co-occurrences and top ten responses is considered, the larger the context window, the higher the overlap for all semantic relations except syntagmatic contiguity. Contrast relations, for instance, are highly

Table 5.13: Correctly predicted primary responses for all lemmatised models (x stands for overlap, - for no overlap)

Stimulus	Correctly predicted primary response	Semantic relation	Context window ±2	Context window ±5	Context window ±10	Context window ±20
бабушка *grandmother*	дедушка *grandfather*	contrast	x	x	x	x
белый *white*	черный *black*	contrast	-	x	x	x
богатый *rich*	бедный *poor*	contrast	-	x	x	x
черный *black*	белый *white*	contrast	x	x	x	x
девочка *girl*	мальчик *boy*	contrast	x	x	x	-
добро *goodness*	зло *evil*	contrast	x	x	x	x
искать *to search*	найти *to find*	sem. contiguity	-	x	x	x
жена *wife*	муж *husband*	contrast	-	x	x	x
женщина *woman*	мужчина *man*	contrast	x	x	x	x
лес *forest*	дерево *tree*	sem. contiguity	-	-	-	x
маленький *small*	большой *big*	contrast	-	x	x	x
мальчик *boy*	девочка *girl*	contrast	x	x	x	x
мать *mother*	отец *father*	contrast	x	x	x	x
молодой *young*	человек *man*	synt. contiguity	x	x	x	x
муж *husband*	жена *wife*	contrast	-	x	x	x
мужчина *man*	женщина *woman*	contrast	x	x	x	x
новый *new*	старый *old*	contrast	-	-	x	x
ночь *night*	день *day*	contrast	x	-	-	-
палец *finger*	рука *hand*	sem. contiguity	-	-	-	x
плохо *bad*	хорошо *good*	contrast	-	-	x	-
радость *joy*	счастье *happiness*	sem. contiguity	x	x	x	x
свет *light*	яркий *bright*	synt. contiguity	x	x	x	x
стыд *shame*	позор *dishonour*	similarity	-	-	-	x
слабый *shame*	сильный	contrast	-	x	x	x
смерть *death*	жизнь *life*	contrast	x	x	x	x
справедливость *justice*	восторжествовать *triumphs*	synt. contiguity	-	x	x	x
умный *intelligent*	человек *man*	synt. contiguity	x	x	x	x
зло *evil*	добро *goodness*	contrast	x	x	x	x

Table 5.14: Semantic relation in overlap between top ten responses and top ten primary co-occurrences for unlemmatised and lemmatised models

		contrast	similarity	sem. contiguity	synt. contiguity
Unlemmatised	Context ±2	21.11% (19)	5.36% (6)	2.60% (10)	25.99% (138)
	Context ±5	31.11% (28)	7.14% (8)	4.16% (16)	25.05% (133)
	Context ±10	36.67% (33)	8.04% (9)	4.42% (17)	23.92% (127)
	Context ±20	41.11% (37)	7.14% (8)	6.23% (24)	20.15% (107)
Lemmatised	Context ±2	33.33% (30)	8.93% (10)	8.57% (33)	23.16% (123)
	Context ±5	50.00% (45)	12.50% (14)	14.55% (56)	22.41% (119)
	Context ±10	52.22% (47)	16.07% (18)	17.14% (66)	20.72% (110)
	Context ±20	56.67% (51)	18.75% (21)	20.00% (77)	19.21% (102)
Russian top ten responses		(90)	(112)	(385)	(531)

represented in the overlap of top ten co-occurrences with top ten responses as well; one can observe that for the model trained on the lemmatised corpus with the context window of ±20 the proportion of overall overlap between co-occurrences and contrast associations amounts to 56.67%, which is more than twice as high than the relative overlap of co-occurrences with responses having any other semantic relations in any model.[72] It may seem surprising that the proportions of similarity and semantic contiguity are very close to one another when the overlap of the top ten is concerned (18.75%; 20%), since it has been expected that words which are similar do not occur together, but that they rather share the same context. A relatively high proportion of overlap based on similarity regards not only hierarchic relations such as любовь-чувство (love-feeling) or co-hyponyms, such as красный-черный (red-black), but also synonyms, for instance машина-автомобиль (car-automobile) or путь-дорога (way-way).

Of course, the choice of literature texts from Lib.ru also had an effect on the overlap between co-occurrences; for instance, certain co-occurrences had become a high score because there were frequently used together with the respective stimulus word in a particular book. Example of such combinations are provided in Table 5.15, in which co-occurrences of the stimulus дядя (uncle) are compared with its most frequent co-occurrences. As it can be observed, the co-occurrences of дядя in the literature corpus tend to be personal names of uncle-figures from different books, out of which дядя Ваня and дядя Петя overlap with human associations while others do not. Other examples in which co-occurrences from corpus contained the names of the book figures which were not as recurring in human associations are бабушка-Иларион, брат-Колько, брат-Регина, брат-Сашка and брат-Петровна.[73]

.

[72]The sum of the Russian responses in Table 5.14 is 1118 and not 1120 because the remaining two associative pairs were connected by phonological similarity, which is not considered further for being extremely rare. For lemmatised models particular overlaps were counted twice, because when lemmatised responses were compared to lemmatised co-occurrences it was not possible to know whether the overlap bases on one or the other word form; for instance, given the stimulus памятник (memorial), the lemmatised co-occurrence Пушкин (Puškin) was compared once to lemmatised response Пушкин originating from (Пушкину[DAT.]) and once from (Пушкин[NOM.]) Hence, it was also not possible to know whether the semantic relation of the overlap refers to syntagmatic contiguity (Пушкину) or semantic contiguity. Therefore, in those cases, I added +1 to the count of semantic contiguity and +1 to the count of syntactic contiguity.

[73]The co-occurrence бабушка-Иларион originates from the "Я, бабушка, Илико и Илларион" written by Nodar Dumbadze (1960); co-occurrences Регина and Петровна refer to the same figure - Регина Петровна from the story "Ночевала тучка золотая" (The Inseparable Twins) from 1987 written by Anatolij Pristavkin; the co-occurrence брат-Колько refers to Koljko from by "Babyj Jar" (1966) by Anatolij Kuznecov; the combination брат-Сашка comes from Сашка in "Poslednij Parad" (2003) by Vjačeslav Degtev. One may assume that these figures are not very often referred to outside of the literature domain and that the content of the books is still not note to a majority of readers such as, for instance Čechov's "Uncle Vanja".

Table 5.15: Responses to the stimulus **дядя** in comparision to its co-occurrences extracted with the model trained on the lemmatised corpus with the context window of ±20 words

Responses	Frequency	Co-occurrences	LLR	Overlap
тетя *aunt*	134	Коля *Kolja*	1308.61	-
Степа *Stepa*	56	тетя *aunt*	1063.85	x
Вася *Vasja*	37	Сандро *Sandro*	929.77	-
родственник *relative*	35	Яша *Jaša*	693.20	-
Ваня *Vanja*	31	Петя *Petja*	611.88	x
родной *dear/close*	23	Володя *Volodja*	582.68	-
мой *my*	15	Миша *Miša*	538.96	-
Петя *Petja*	12	Чик *Čik*	448.77	-
друг *friend*	9	Ваня *Vanja*	429.98	x
мужчина *man*	8	Гриша *Griša*	416.05	-

5.4 Discussion

According to the results presented in this work, a little less than a quarter of human associations consists of the words occurring most frequently together within the context of 40 words (context window of ±20 words). With an overlap of 22.32%, the proportion of co-occurrences which correspond to associations found in this study is higher the one observed in Rapp & Wettler (1991), as well as the one presented by Griffiths & Steyvers (2002), which was based on more sophisticated, second-order statistics models such as Latent Semantic Analysis (overlap: 11.16%) and topic model (overlap: 12.86%). Although the overlap of primary co-occurrences and primary responses did not overcome 29% measured by Wettler *et al.* (2005), with 70% of co-occurrences present within any human response extracted with the context window of ±10 words, the overlap found in this work was higher than the one measured by Wettler and colleagues, which amounted to 64%. Comparing the results with those made by other researches gives an idea of the overlap that is to be expected when predicting human responses with methods of distributional semantics and reassures that the overlap observed within this work is fairly satisfying. However, claiming that the results found within the parameters used in this study are more adequate than other models used for the same purpose would be an overstatement, since the comparison between results of models which were trained on different corpora, different languages and which make use of different lists of stimuli does not hold. The parameters which have been observed to extract co-occurrences with the highest degree of overlap (lemmatised corpus, context window of ±20 words) are therefore not to be consid-

ered as general formula assuring the highest overlap between associations and co-occurrences, but as the one which proved to be the most adequate for the Russian associations presented in Slavic Association Dictionary. Accordingly, the parameter combination found to be best adequate for Russian association would possibly not be as well adapted for other languages, since, as it has been shown in the analysis of Slavic associations, different languages display different distributions of semantic relations within associations even when the same list of stimuli is presented. Moreover, it is not to exclude that even the distribution of semantic relations for the same language would vary with the repetition of the associative experiment with other subjects (or maybe even with the same subjects).

The more context words were considered by the co-occurrence extraction, the closer were the distributions of semantic relations within co-occurrences and human responses. The distribution of semantic relations within co-occurrences depended on the size of the context window: the smaller the window size, the higher the proportion of syntagmatic contiguity could be observed, the wider the window size, the higher the proportion of all the other semantic relations. What nouns and adjectives are concerned, it has been shown that the proportion of homogeneous (paradigmatic) and, respectively, heterogeneous (syntagmatic) pairs was dominant for both data sets. Hence, it may be stated that co-occurrences certainly have a potential to overlap with human associations, since the relations present in human data are also present in a similar proportion within co-occurrence data. This potential is, however, not necessarily realized for each semantic relation in equal measure. As it has been shown in the analysis of the overlap, although the majority of co-occurrences was based on syntagmatic contiguity, the overlap between primary responses and co-occurrences consisted almost exclusively of contrast relations: 16 out of 21 opposite primary co-occurrences were present in primary responses. Accordingly, the proportion of overlap seems to depend on the proportion of words which might have a contrast in the list of stimuli: the more stimuli with a potential contrast response are present, the higher overlap can be expected. The great proportion of contrast co-occurrences confirms the findings of Charles & Miller (1989) and Fellbaum (1995), who noted that opposite words tend to appear in the same context more often than chance.

The presence of verbs in the list of stimuli also seems to correlate with the proportion of overlap what primary co-occurrences as well as top ten co-occurrences are concerned. The reason for a lower overlap seems to be the fact that co-occurrences of verb stimuli tend to

be functional words such as particles and pronouns which people generally do not associate. Pronouns and other functional part-of-speech tags are more frequently found in co-occurrences than in human responses: as observed in the analysis of all Slavic associations, to content word stimuli, such as nouns, adjectives and verbs, people most commonly associate other content words and not functional words, although the latter may be more commonly encountered in the same context with the stimulus than content words. Consequently, it is legitimate to hypothesise that content word associations reflect only content word co-occurrences.

According to the results of this analysis, when considering the similarity measures such as LSA for investigating the overlap between extracted *related words* and human associations, one must take into consideration that associations often consist of words which may be found in the same context, but which are not necessarily similar to each other. Hence, a large proportion of syntagmatic relations would possibly be more difficult to capture than in the case of association prediction with association measures.

As stated by Rapp & Zock, deriving associations from textual co-occurrences pre-supposes that the human brain is also doing this (2014:4). Hence, ignoring function words from predicted associations pre-supposes that human brain is doing the same. With the aim of finding the parameter combination that yields the highest overlap with human responses, I experimented with ignoring conjunctions and adpositions from the resulting co-occurrences, which finally proved to slightly increase the proportion of the overlap. To be completely in line with the contiguity hypothesis, I must recall that by doing so, I assumed that human mind is also ignoring these parts-of-speech in case content words stimuli are presented. However, even when strictly remaining by the traditional hypothesis, without part-of-speech constraints, the change in the overlap proportion of primary co-occurrences was only slightly lower (21.43%). Instead, the lemmatisation proved to be much more important factor for achieving the high overlap, as it was expected for a highly morphological language such as Russian.

Although a correlation between co-occurrences and human associations has been observed, with the observed results it is not possible to determine whether the textual co-occurrences influence associations or whether association influence textual co-occurrences. It might certainly be assumed that there is a mutual influence between the two, since our thoughts may influence what we say just as well as the words we say may influence our thoughts, and thus, a creation of an associative link between words in our mind.

6 Conclusion

In the present study, I investigated the overlap between human association and co-occurrences in text by comparing Russian association norms from Slavic Association Dictionary with co-occurrences extracted from a Russian corpus consisting of literary texts. In order to study the overlap between the two, I provided a deep analysis of associations and co-occurrences regarding the distribution of their parts-of-speech, as well as of the distribution of the semantic relations such as contrast, similarity and syntagmatic and semantic contiguity. The highest overlap between associations and co-occurrences as well as between semantic properties of the two has been observed by the results of the model trained on the lemmatised corpus with the context size of ± 20 words, which also proved to cover a higher amount of semantic relations present in human associations than the models trained on smaller context windows. It has been shown that the proportion of the overlap depends not only on the association measure, corpus, context window and lemmatisation, but also on the content of association norms to which co-occurrences are compared to, as well as on the language(s) in which the associative experiment has been conducted. Since association norms of different Slavic languages had shown different distributions of semantic relations, in further studies it would be interesting to investigate to what extent the distribution of semantic relations within co-occurrences is language-independent.

There are numerous other directions in which the investigation of the correlation between associations and co-occurrences in language could be extended. For instance, exploring whether there is a correlation between the variance of human responses and the proportion of overlap between co-occurrences and associations would be an intriguing research subject. The assumption regarding the variance could be that certain stimuli words, to which participants agreed the most, are also those by which a higher overlap could be expected. Another interesting subject of research would be to investigate whether associations and co-occurrences tend to precede or succeed the target word.

7 Bibliography

Aitchison, J. (2003). *Words in the Mind: an Introduction to the Mental Lexicon.* Oxford.

Anstatt T. (2008). Wer "dunkel" hört, muss nicht "hell" sagen: Wortassoziationen in slavischen und germanischen Sprachen. In: Kosta, Peter / Weiss, Daniel (eds.): *Slavistische Linguistik 2006/2007. Referate des XXXII. und XXXIII. Konstanzer Slavistischen Arbeitstreffens* (= Slavistische Beiträge 464). München, 11-34.

Blank, A. (1997). *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen.* Tübingen.

Brown, R. & Berko, J. (1960). Word association and the acquisition of grammar. In: *Child Development* 31, 1–14.

Bruni, E., Tran, N. K. & Baroni, M. (2013). Multimodal distributional semantics. In: *Journal of Articifial Intelligence Research* 48, 1–47.

Charles, W. & Miller, G. (1989). Contexts of antonymous adjectives. In: *Applied Psycholinguistics* 10. 357-75.

Church, K. W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. In: *Computational Linguistics* 16(1). 22-29.

Clark, H. H. (1970). Word associations and linguistic theory. In: J. Lyons (ed.): *New horizons in linguistics.* Harmondsworth, 271-286.

Collins, A. M. & Loftus, E. F. (1975). A spreading- activation theory of semantic processing. In: *Psychological Review* 82, 407-428.

Collins, A. M. & Quillian, M. R. (1969). Retrieval time from semantic memory. In: *Journal of Verbal Learning and Verbal Behavior* 8, 240-247.

Cramer, P. (1968). *Word association.* New York.

Cruse, D. A. (1986). *Lexical Semantics.* Cambridge.

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques.* Ph.D. thesis, Université Paris 7.

De Deyne, S. & Storms, G. (2008). Word associations: network and semantic properties. In: *Behavior Research Methods* 40(1), 213–231.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990). Indexing by latent semantic analysis. In: *Journal of the American Society for Information Science* 41, 391-407.

Deese, J. (1962). Form class and the determinants of association. In: *Journal of Verbal Learning and Verbal Behavior* 1, 79–84.

Deese, J. (1965). *The Structure of Association in Language and Thought.* Baltimore.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. In: *Computational Linguistics* 19 (1), 61–74.

Entwisle, D. R. (1966). *The word associations of young children.* Baltimore.

Ervin, S. M. (1961). Changes with age in the verbal determinants of word association. In: *American Journal of Psychology* 74, 361-372.

Evert, S. (2005). *The statistics of word co-occurrences: Words pairs and collocations.* Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Evert, S. & Krenn B. (2001). Methods for the qualitative evaluation of lexical association measures. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, Toulouse (France), 188–195.

Fano, R. (1961). *Transmission of Information: A Statistical Theory of Communications.* Cambridge.

Fellbaum, C. (1995). Co-occurrence and Antonymy. In: *International Journal of Lexicography* 8.4, 281-303.

Fellbaum, C. (1998). *Wordnet: an Electronic Lexical Database.* Cambridge.

Fernando, S. & Stevenson, M. (2008). A semantic similarity approach to paraphrase detection. In: *Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*, 45–52.

Filippovič, Ju.N., Čerkasova, G.A. & Del'ft, D. (2002). *Associacii informacionnyx texnologij: eksperiment na russkom i francuzskom jazykax.* (s.l.).

Galton, F. (1880). Psychometric experiments. In: *Brain* 2, 149-162.

Griffiths, T. L. & Steyvers, M. (2003). Prediction and Semantic Association. In: *Advances in Neural Information Processing Systems* 15, 11–18.

Harris, Z. (1954). Distributional structure. In: *Word* 10(23), 146-162.

Huang, A. (2008). Similarity measures for text document clustering. In: *Proceedings of the New Zealand Computer Science Research Student Conference (NZCSRSC)*, 49–56.

James, W. (1890). *The principles of psychology.* New York.

Jenkins, J. J. (1970). The 1952 Minnesota word association norms. In: L. Postman, G. Keppel (eds.): *Norms of word association.* New York, 1-38.

Jones, E. (1964). *The Life and Works of Sigmund Freud.* Harmondsworth.

Jones, M. N. & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. In: *Psychological Review* 114(1), 1-37.

Jones, S. (2002). *Antonymy. A corpus-based perspective.* London/New York.

Jung, C. G. (1918). *Studies in word-association: experiments in the diagnosis of psychopathological conditions carried out at the Psychiatric Clinic of the University of Zurich.* London.

Karaulov, Ju. N., Sorokin, Ju. A., Tarasov, E. F., Ufimceva, N. V. & Čerkasova, G. A. (1994ff). *Russkij associativnyj slovar'. Associativnyj tezaurus sovremennogo russkogo jazyka.* Moskva.

Kent, G. H. & Rosanoff, A. J. (1910). A study of association in insanity. In: *American Journal of Insanity* 67, 37-96.

Kiss, G., Armstrong, C., Milroy, R. & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In: A. Aitken, R. Beiley and N. Hamilton-Smith (eds.): *The Computer and Literary Studies.* Edinburgh, 153-165.

Kurcz, I. (1967). Polskie normy powszechności skojarzeń swobodnych na 100 słów z listy Kent-Rosanoff'a. In: *Studia Psychologiczne* 8, 122-255.

Lakoff, G. (1993). The Contemporary Theory of Metaphor. In A. Ortony (ed.): *Metaphor and Thought*. Cambridge, 2-202.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. In: *Psychological Bulletin* 104, 211-240.

Lehmann, C. (1982). Directions for interlinear morphemic translations. In: *Folia Linguistica* 16, 199-224.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. In: *Italian Journal of Linguistics 20(1)*, 1-30.

Leont'ev, A. A. (1977). *Slovar' associativnych norm russkogo jazyka*. Moskva.

Lezius, W. (1999). Automatische Extrahierung idiomatischer Bigramme aus Textkorpora. In: *Tagungsband des 34. Linguistischen Kolloquiums*, Germersheim.

Lobanova, A. (2012). *The Anatomy of Antonymy: a Corpus-driven Approach*. Dissertation. University of Groningen.

Lü, Y. and M. Zhou (2004). Collocation Translation Acquisition Using Monolingual Corpora. In: *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL 2004)*. Barcelona, 167–174.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. In: *Behavior Research Methods, Instruments, & Computers* 28(2), 203–208.

Lyons, J. (1977). *Semantics*. Cambridge.

Mikolov T., Chen K., Corrado G., Dean J. (2013a). Efficient Estimation of Word Representations in Vector Space. Available at: http://arxiv.org/abs/1301.3781 (Last accessed: 07.12.2015).

Mikolov, T., Yih, W. & Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In: *Proceedings of Human Language Technologies: Conference of the*

North American Chapter of the Association of Computational Linguistics (NAACL-HLT). Georgia (USA), 746–751.

Miller, K. M. (1970). Free-association responses of English and Australian students to 100 words from the Kent-Rosanoff association test. In: L. Postman & G. Keppel (eds.): *Norms of Word Association*. New York, 39-52.

Nelson, D. L., McEvoy, C. L. & Dennis, S. (2000). What is free association and what does it measure? In: *Memory & Cognition* 28, 887–899.

Nelson, D. L., McEvoy, C. L. & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. In: *Behavior Research Methods, Instruments, & Computers* 36(3), 402–407.

Nelson, K. (1977). The syntagmatic-paradigmatic shift revisited: a review of research and theory. In: *Psychological Bulletin* 84(1), 93–116.

Orliac, B. & Dillinger, M. (2003). Collocation Extraction for Machine Translation. In: P*roceedings of Machine Translation Summit IX*. New Orleans, 292–298.

Palermo, D. S. (1971). Characteristics of word association responses obtained from children in grades one through four. In: *Developmental Psychology* 5(1), 118-123.

Palermo, D. S. & Jenkins, J. J. (1964). *Word Association Norms: Grade School through College*. Minneapolis.

Panchenko A., Loukachevitch N. V., Ustalov D., Paperno D., Meyer C. M. & Konstantinova N. (2015). RUSSE: The First Workshop on Russian Semantic Similarity. In: P*roceedings of the Dialogue 2015 conference*. Moscow, xx-yy.

Paperno, D., Marelli, M., Tentori, K. & Baroni, M. (2014). Corpus-based estimates of word association predict biases in judgement of word co-occurrence likelihood. In: *Cognitive Psychology* 74, 66–83.

Pereira, F., Tishby, N. & Lee, L. (1993). Distributional clustering of English words. In: *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*. Columbus, 183–190.

Petrov, S., Das, D. & McDonald, R. (2011). A universal part-of-speech tagset. Available at http://arxiv.org/pdf/1104.2086.pdf (Last accessed: 07.12.2015).

Postman, L. & Keppel, G. (eds.) (1970): *Norms of Word Association.* New York/London.

Raible, W. (1981). Von der Allgegenwart des Gegensinns. In: *Zeitschrift für Romanische Philologie* 97, 1-40.

Rapp, R. (2002). The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In: *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei (Taiwan), 1–7.

Rapp, R., Wettler, M. (1991). Prediction of free word associations based on Hebbian learning. In: *Proceedings of the (IEEE and INNS) International Joint Conference on Neural Networks.* Singapore, 25–29.

Rapp, R. & Wettler, M. (1993). Computation of word associations based on the co-occurrence of words in large corpora. In: *Proceedings of the Workshop on Very Large Corpora.* Columbus (OH), 84–93.

Rapp, R. & Zock, M. (2014). The CogALex-IV Shared Task on the Lexical Access Problem. In: *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex).* Dublin, 1–14.

Russel, W. A. (1970). The complete German language norms for responses to 100 words from the Kent-Rosanoff Word Association Test. In: Postman, L./Keppel, G. (eds.): *Norms of word association.* New York, 53-94.

Sahlgren, M. (2006). *The Word-Space Model Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* Ph.D. Thesis, Stockholm University.

Saussure, F. (1916/1983). *Course in general Linguistics.* Duckworth.

Schulte im Walde, S. & Melinger, A. (2008). An in-depth look into the co-occurrence distribution of semantic associates. In: *Rivista Di Linguistica*, 20(1), 87–123.

Schwartz, B. & Reisberg, D. (1991). *Learning and memory.* New York.

Seretan, V. (2010): *Syntax-Based Collocation Extraction. Text, Speech and Language Technology.* Dordrecht.

Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A. & Divjak, D. (2008). Designing and evaluating Russian Tagsets, In: *Proceedings of the Sixth Language Resource and Evaluation Conference (LREC).* Marrakech, 279-285.

Simov, K., Osenova, P. & Slavcheva, M. (2004). BTB-TR03: BulTreeBank Morphosyntactic Tagset. BTB-TS version 2.0. Available at: http://www.bultreebank.org/TechRep/BTB-TR03.pdf. (Last accessed: 06.12.2015).

Sorabji, R. (2004). *Aristotle on memory.* Chicago.

Steyvers, M. & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. In: *Cognitive Science*, 29(1), 41–78.

Suprun, A. E. (1983). *Leksičeskaja tipologija slavjanskich jazykov.* Minsk.

Thumb, A. & Marbe, K. (1901). *Experimentelle Untersuchungen über die1 psychologischen Grundlagen der sprachlichen Analogiebildung.* Leipzig.

Turney, P. D. & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. In: *ACM Transactions on Information Systems* 21(4), 315–346.

Ufimceva, N. V. G. A. Čerkasova, Ju. N. Karaulov, E. F., Tarasov, M. (2004). *Slavjanskij associativnyj slovar': russkij, belorusskij, bolgarskij, ukrainskij.* Moskva.

Wandmacher, T., Ovchinnikova, E. & Alexandrov, T. (2008). Does latent semantic analysis reflect human associations. In: *Proceedings of the Lexical Semantics Workshop at ESSLLI.* Hamburg, 63-70.

Wettler, M., Rapp, R. & Sedlmeier, P. (2005). Free Word Associations Correspond to Contiguities Between Words in Texts. In: *Journal of Quantitative Linguistics*, 12(2-3), 111–122.

8 Attachment

Table 8.1: Russian stimuli from Slavic Association Dictionary

бабушка	*grandmother*	женщина	*woman*	родина	*homeland*
белый	*white*	жизнь	*life*	родной	*close/dear*
Бог	*God*	жить	*to live*	рот	*mouth*
богатый	*rich*	красивый	*beautiful*	руки	*hands*
больной	*ill*	красный	*red*	свет	*light*
большой	*big*	кричать	*to yell*	старый	*old*
брат	*brother*	лес	*forest*	стол	*table*
быстро	*fast*	лицо	*face*	стыд	*shame*
человек	*man*	любовь	*love*	счастье	*hapiness*
черный	*black*	маленький	*small*	свободный	*free*
чистый	*clean*	мальчик	*boy*	семья	*family*
дверь	*door*	мать	*mother*	сила	*force*
дочь	*daugther*	машина	*car*	слабый	*weak*
друг	*friend*	много	*a lot*	слово	*word*
думать	*think*	молодой	*young*	смерть	*death*
дурак	*idiot*	муж	*husband*	справедливость	*justice*
душа	*soul*	мужчина	*man*	терять	*to lose*
дядя	*oncle*	надеяться	*to hope*	веселый	*joyful*
девочка	*girl*	народ	*nation*	вспоминать	*to remember*
дело	*action*	начало	*beggining*	встреча	*meeting*
день	*day*	ненавидеть	*to hate*	ветер	*wind*
деньги	*money*	новый	*new*	вечер	*evening*
деревня	*village*	ночь	*night*	вечность	*eternity*
добро	*goodness*	обещать	*to promise*	вместе	*together*
дом	*home*	обман	*fraud*	вода	*water*
есть	*to eat*	огонь	*fire*	война	*war*
глаза	*eye*	палец	*finger*	враг	*enemy*
глупый	*stupid*	памятник	*memorial*	время	*time*
говорить	*to speak*	пить	*to drink*	хлеб	*bread*
голова	*head*	плохо	*bad*	ходить	*go*
голос	*voice*	помогать	*to help*	хорошо	*good*
гора	*muntain*	путь	*way*	хотеть	*to want*
город	*town*	работа	*work*	умный	*intellingent*
гость	*guest*	радость	*joy*	успеть	*to succeed*
искать	*to search*	разговор	*conversation*	утро	*morning*
жадный	*avid*	ребенок	*child*	зеленый	*green*
жена	*wife*	река	*river*	земля	*earth*
				зло	*evil*

Table 8.2: Distribution of parts-of-speech for noun, adjective and verb stimuli in Russian, Belarusain, Bulgarian and Ukrainian **primary responses** (absolute values)

		NOUN	ADJ	VERB	ADV	NUM	PRON
NOUN (total: 71)	ru	52	13	4	0	1	1
	be	55	16	0	0	0	0
	bg	62	7	1	0	1	0
	uk	24	47	0	1	0	0
ADJ (total: 20)	ru	13	7	0	0	0	0
	be	18	2	0	0	0	0
	bg	9	11	0	0	0	0
	uk	20	0	0	0	0	0
VERB (total: 16)	ru	3	0	10	3	0	0
	be	8	0	5	3	0	0
	bg	5	0	10	1	0	0
	uk	7	0	5	4	0	0

Table 8.3: Distribution of parts-of-speech for noun, adjective and verb stimuli in Russian, Belarusian, Bulgarian and Ukrainian **ten most frequent responses** (absolute values)

		NOUN	ADJ	VERB	ADV	NUM	PRON
NOUN (total: 710)	ru	435	194	35	19	5	21
	be	484	183	18	9	0	15
	bg	547	114	21	17	4	7
	uk	350	312	29	14	3	12
ADJ (total: 200)	ru	152	45	0	0	0	3
	be	147	50	0	1	0	2
	bg	115	78	2	2	1	2
	uk	172	27	0	0	0	1
VERB (total: 160)	ru	53	1	69	25	5	7
	be	86	3	51	15	4	1
	bg	76	2	46	17	8	11
	uk	85	0	37	32	3	3

Table 8.4: Observed and predicted primary responses for the lemmatised model with context window of ±20 (excluding prepositions and conjunctions)

Stimulus	Primary response	Nr. of subjects with observed primary response	Predicted response	LLR	Nr. of subjects with predicted primary response	Nr. of all subjects	Presence of predicted primary response in any response	Primary response equal to predicted primary response
добро	зло	221	зло	2767.75	221	590	x	x
зло	добро	167	добро	2767.75	167	591	x	x
муж	жена	167	жена	1693.3	167	592	x	x
молодой	человек	163	человек	1965.79	163	591	x	x
мужчина	женщина	156	женщина	3344.44	156	594	x	x
плохо	хорошо	152	хорошо	222.21	152	590	x	x
жена	муж	129	муж	1693.3	129	593	x	x
бабушка	дедушка	128	дедушка	1552.31	128	592	x	x
мальчик	девочка	124	девочка	947.89	124	590	x	x
черный	белый	121	белый	1939.26	121	592	x	x
белый	черный	119	черный	1939.26	119	592	x	x
слабый	сильный	117	сильный	390.78	117	590	x	x
умный	человек	113	человек	458.11	113	592	x	x
мать	отец	111	отец	3343.77	111	588	x	x
богатый	бедный	108	бедный	389.62	108	592	x	x
девочка	мальчик	101	мальчик	947.89	101	592	x	x
новый	старый	95	старый	704.23	95	592	x	x
женщина	мужчина	89	мужчина	3344.44	89	593	x	x
смерть	жизнь	88	жизнь	961.18	88	592	x	x
маленький	большой	82	большой	576.59	82	593	x	x
радость	счастье	79	счастье	252.34	79	591	x	x
искать	найти	75	найти	931.94	75	592	x	x
свет	яркий	71	яркий	948.4	71	591	x	x
дверь	открытый	58	открыть	3146.52	58	592	x	-
ветер	сильный	68	дуть	1427.67	53	589	x	-
земля	круглый	79	небо	1202.01	50	591	x	-
вечер	теплый	52	утро	1197.18	49	592	x	-
палец	рука	141	указательный	1804.23	47	593	x	-

враг	друг	102	народ	1374.35	40	592	x	-
справедливость	восторжествовать	39	восторжествовать	156.41	39	592	x	x
дочь	сын	108	мать	718.73	39	593	x	-
терять	находить	85	время	530.78	37	590	x	-
деревня	сесть	69	город	193.66	36	593	x	-
зеленый	цвет	80	красный	459.49	32	594	x	-
огонь	вода	106	гореть	927.21	25	590	x	-
старый	дед	109	новый	704.23	23	591	x	-
помогать	мама	45	друг	143.7	22	590	x	-
родной	язык	195	брат	387.09	20	592	x	-
есть	вкусно	39	пить	788.19	20	589	x	-
брат	сестра	124	старший	1703.48	20	511	x	-
стол	стул	118	письменный	3313.27	16	592	x	-
река	вода	75	берег	2938.39	15	591	x	-
больной	человек	112	врач	389.83	14	592	x	-
ночь	день	105	спать	2414.26	12	592	x	-
утро	вечер	90	ночь	1197.65	12	591	x	-
гора	высокий	72	вершина	945.37	11	592	x	-
пить	вода	107	чай	3908.92	11	593	x	-
глаз	голубой	83	смотреть	2681.02	10	592	x	-
вместе	всегда	25	мы	537.28	8	593	x	-
стыд	позор	70	чувство	210.6	7	590	x	-
кричать	громко	143	ура	336.85	6	592	x	-
лес	дерево	55	опушка	814.89	6	594	x	-
много	деньги	121	год	1069.83	6	593	x	-
дурак	умный	78	ты	499.05	6	592	x	-
дядя	тетя	134	коля	1300.27	5	592	x	-
чистый	грязный	88	вода	336.25	5	592	x	-
слово	дело	59	честный	1417.89	4	591	x	-
красивый	человек	57	женщина	684.45	4	593	x	-
человек	животное	25	молодой	1965.79	3	588	x	-
жить	хорошо	71	дом	924.43	3	589	x	-
родина	мать	221	наш	332.51	3	592	x	-
город	курск	96	улица	894.31	3	594	x	-
дом	родной	57	улица	1352.45	2	590	x	-
веселый	человек	75	смеяться	162.9	2	586	x	-
любовь	счастье	63	любить	575.96	2	594	x	-
лицо	красивый	68	выражение	2927.35	2	588	x	-

хорошо	плохо	181	знать	784.46	2	592	x	-
памятник	пушкин	49	поставить	186.63	2	589	x	-
вспоминать	прошлое	68	часто	182.41	1	593	x	-
говорить	молчать	43	я	4399.31	1	593	x	-
глупый	человек	83	ты	184.7	1	594	x	-
деньги	много	45	купить	864.89	1	593	x	-
жизнь	смерть	99	весь	2211.52	1	593	x	-
хотеть	быть	51	я	6139.53	1	594	x	-
встреча	радость	51	состояться	221.31	1	590	x	-
война	мир	102	гражданский	2056.04	1	590	x	-
думать	мыслить	27	я	3580.06	1	593	x	-
обман	ложа	103	оптический	75.314	1	477	x	-
вода	чистый	63	берег	1789.21	0	594	-	-
начало	конец	227	самый	971.03	0	591	-	-
надеяться	ждать	94	вы	256.75	0	592	-	-
красный	цвет	80	армия	2631.53	0	593	-	-
рот	зуб	56	изо	1964.68	0	593	-	-
бог	быть	42	слава	4154.69	0	592	-	-
счастье	быть	74	попытать	260.06	0	592	-	-
время	деньги	126	тот	3412.89	0	508	-	-
хлеб	еда	48	кусок	1544.58	0	593	-	-
большой	дом	85	не	659.45	0	593	-	-
народ	толпа	84	враг	1374.35	0	593	-	-
рука	нога	109	махнуть	3751.92	0	593	-	-
дело	время	45	самый	4414.14	0	594	-	-
ненавидеть	любить	96	я	258.91	0	593	-	-
семья	мой	57	жить	484.17	0	593	-	-
ходить	бегать	57	слух	365.31	0	592	-	-
голова	болеть	67	покачать	4596.13	0	590	-	-
день	ночь	175	следующий	2571.2	0	589	-	-
работа	деньги	38	научный	324.09	0	590	-	-
успеть	опоздать	56	не	769.85	0	592	-	-
обещать	выполнять	33	я	139.95	0	591	-	-
вечность	жизнь	25	целый	159.28	0	508	-	-
ребенок	маленький	93	женщина	286.2	0	593	-	-
душа	тело	52	глубина	709.97	0	593	-	-
друг	враг	82	они	2754.49	0	593	-	-
голос	звук	40	услышать	2007.12	0	591	-	-
машина	время	66	шофер	902.2	0	508	-	-

гость	нежданный	67	пригласить	366.51	0	594	-	-
быстро	медленно	105	довольно	241.38	0	589	-	-
сила	воля	40	изо	2491.6	0	591	-	-
свободный	человек	141	время	412.28	0	590	-	-
путь	дорога	196	обратный	1405.72	0	593	-	-
жадный	человек	84	глоток	57.99	0	590	-	-
разговор	беседа	52	вести	510.2	0	590	-	-

Figure 8.1: Distribution of parts-of-speech for noun, adjective and verb stimuli for ten most frequent responses and ten most frequent co-occurrences extracted with model trained on lemmatised corpus with the context window of ±20

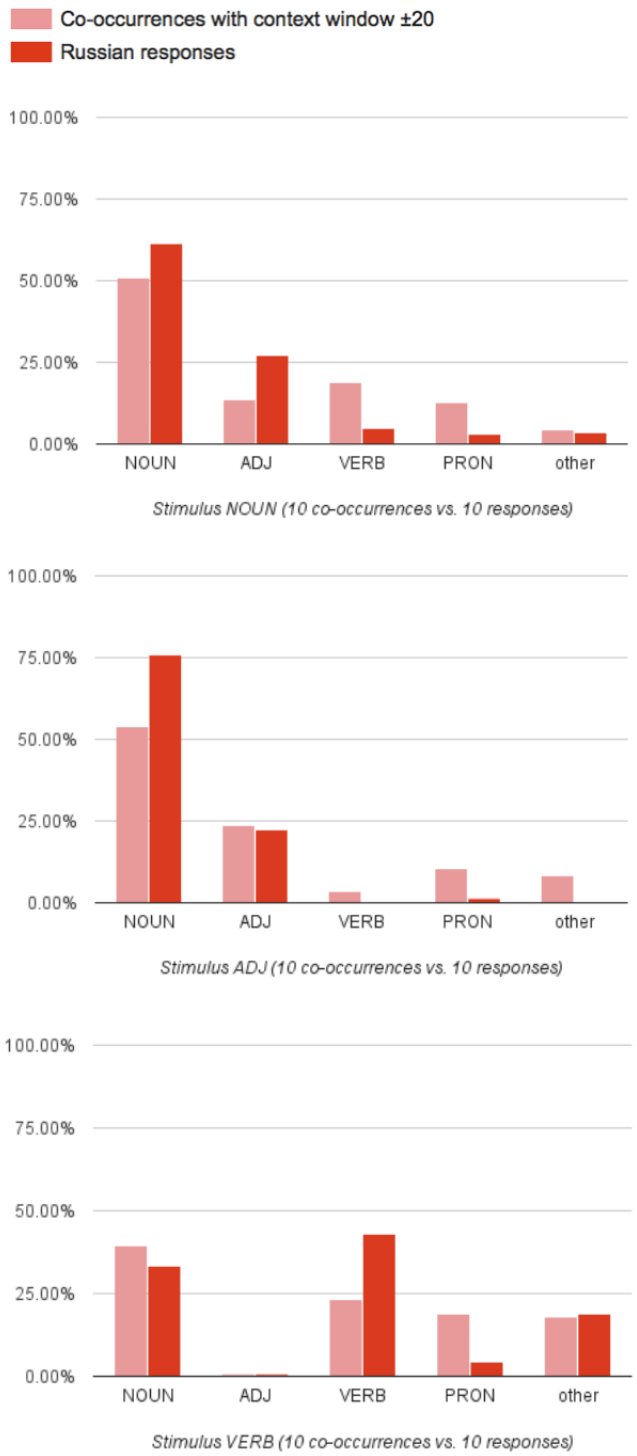
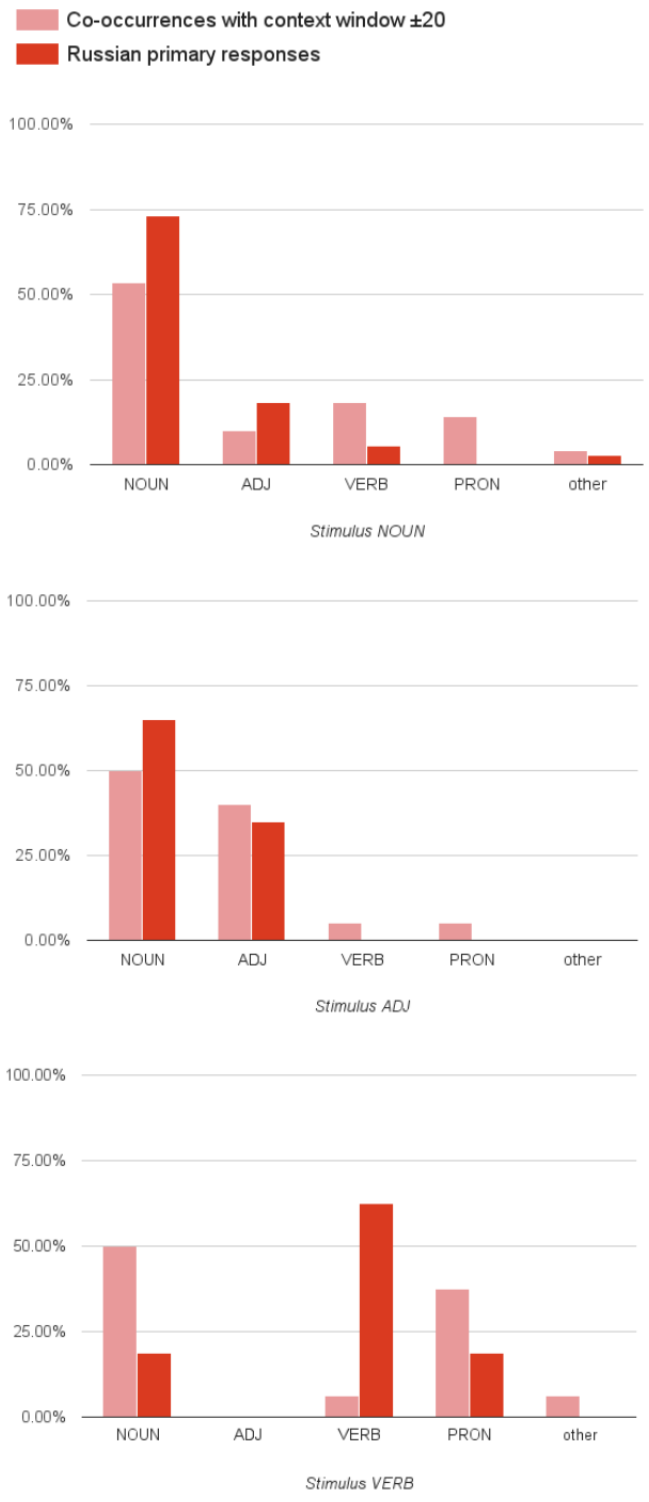


Figure 8.2: Distribution of parts-of-speech for noun, adjective and verb stimuli for primary responses and primary co-occurrences extracted with model trained on lemmatised corpus with the context window of ±20



List of Figures

List of Tables