Measuring Similarity of Fiction Texts Based on Distributional Semantic Models (Case Study of the Russian Original Text and English Translations of M.Bulgakov's Novel "The Master and Margarita")

Ekaterina V. Tretyak

***Abstract*— The paper deals with the application of distributional semantic methods to the task of measuring similarity between several translations of the original text. In particular, Word2Vec neural network toolkit is employed for comparison between two translations. Moreover, in terms of the theory of translation, descriptions of transformations for paraphrasing, which are also used for testing plagiarism detection methods, suit the task of comparing translations. Experiments discussed in this paper are carried out for the Russian original and English translations of M. Bulgakov's novel "The Master and Margarita". In the paper, the above mentioned approaches are combined to contrast the translation by M. Glenny (1967) with one by R. Pevear and L. Volokhonsky (1997). Hypothesis that parallel translations can be treated as paraphrases obtained as a result of transformations is under consideration. The paper contains detailed quantitative analysis of the data obtained regarding the similarity between two translations of fiction text as well as discussion of particular contexts.**

***Keywords*— processing of fiction texts, translation, paraphrase transformations, word embeddings, distributional semantic collocation extraction**

I. Introduction

In modern computational linguistics a group of tasks dealing with automatic processing of translated texts is extremely popular and in great demand. They include creation of parallel and comparable corpora, their subsequent alignment, extraction of translation equivalents, measuring similarity of original texts and their translations, evaluating correctness of text translation into a foreign language. These tasks require different approaches than machine translation.

The aim of this study is to measure the similarity between two translations of fiction text from Russian into English based on the techniques and models of distributional semantics. In addition, in this research we identify lexical translation correspondences for lemmas and phrases.

The following tasks are solved in order to achieve the study objective:

1) creating an aligned parallel corpus of the text of Mikhail Bulgakov's novel "The Master and Margarita" and its English translations by Michael Glenny (1967) and Richard Pevear and Larissa Volokhonsky (1997);
2) extracting a sample of translated sentences that correspond the requirements of the experiment and performing operations on sentence vectors in a multi-dimensional vector space;
3) measuring similarity of the pairs of two English translations in terms of cosine measure;
4) classifying transformations, that is, synonymic modifications, based on linguistic analysis of the pairs of the original and translations;
5) analysing the features of each pair of translations in terms of paraphrase typology and the transformations' weights;
6) establishing an appropriate approach for translation equivalents extraction for lemmas and phrases and employ this method in the experiments.

The rise of research interest in the application of techniques and models of distributional semantics for various linguistic problems makes our project extremely actual. To reach our goal, we based our experiments on the theories and methods developed in different fields of knowledge and worked out our solution combining approaches accepted in translation studies, corpus linguistics and distributional semantics.

II. Related Works

Analysis of current research shows that the problem of automatic extraction of translation equivalents can be approached from various sides and a significant number of academic papers in different languages cover this problem. As a rule, it comes down to the extraction of variants of collocations.

In this respect, most strategies use statistical association measures for n-gram windows of different sizes [1], [2]. Other approaches involve syntactic dependence [3] to better

Статья получена 16 декабря 2019. The paper develops the ideas of distributional semantics discussed in the talks of Computational Linguistics and Digital Ontologies Workshop, Internet and Modern Society Conference, IMS 2019.

E. Tretyak is with the Saint Petersburg State University, Saint Petersburg, Russia (e-mail: evtretyak1999@gmail.com).

recognize phrases that occur in the actual syntactic relationship.

Recently, the accessibility of improved parsers has allowed researchers to combine automatically obtained syntactic information with statistical techniques to extract collocations with greater precision [4], [5]. Some researchers focus on extraction of collocations based not only on parallel corpora, but also on semantic classification of the obtained phrases in order to make them more beneficial for applications related to natural language processing [6].

A number of authors also employ parallel corpora so that they can identify translation equivalents, particularly for noun phrases in English and French (e.g. [7]). The proposed method consists of applying EM algorithm (Expectation-maximization algorithm) in previously extracted monolingual collocations. In the same way, in [8] authors obtain Japanese-English equivalents – collocations by calculating their MI scores (Mutual Information scores) and taking into account their frequency and state in the aligned corpora.

In [9], the authors working with parallel corpora extract Chinese and English n-grams from the aligned sentences by calculating their log-likehood ratio. Then the algorithm of comparable connection is employed to find out whether each bilingual pair corresponds to a translation equivalent.

The study [10], although it does not focus on collocations, uses distributional semantic methods for a bilingual dictionary development from the comparable corpora. This approach takes into account the fact, that in this type of resource the state and frequency of an input word cannot be corresponded to those of an output word, and translations of the input words may not be available in the output document.

It should be noted that the compositional distributional semantic methods assure successful performance of the task of extraction of collocations. At present, the attention of researchers is focused on working with distributional multilingual word embeddings. Measuring semantic association includes paradigmatic relations between lexical items. In [11] three distributional methods of modeling the choice of co-occurrence are compared. The first of them calculates the semantic similarity, based on the cosine similarity measure, the second is based on the idea of T. Mikolov [12], which is that language regularities can be captured with the help of distributional vector space models of words, and the last one involves the use of other linguistic resources such as WordNet.

Researchers from the Institute of Informatics Problems of the Russian Academy of Sciences (IPI RAN), describe the model they have developed for extraction of translation equivalents from parallel texts of academic patents in Russian and French, which were obtained from the website of the European Patent Office EPO [13]. The authors used distributional semantic methods, which were worked out in course of the experiment. The authors implemented the vector space model of the corpus in question, similarity of words was calculated by means of cosine measure for word vectors. Thus, the closest word of the French language was chosen as the most relevant translation equivalent for each word of the Russian language, given that the greater the cosine value, the stronger the association between the words of the two languages.

III. Experimental Setup

3.1 Linguistic data

We have chosen a parallel corpus, which includes the Russian original of the novel by M.A. Bulgakov "The Master and Margarita" and its two English translations, as the empirical basis of the experiments. The first of these is made by M. Glenny (1967), and the second one by R. Pevear and L. Volokhonsky (1997). The chosen linguistic data seems to be extremely appropriate for our study for several reasons.

First of all, the world-famous novel by M. A. Bulgakov, even almost a hundred years later after its writing, attracts much attention of Russian readers of various ages and is the object of book, theater and film adaptations. However, year by year this work, which depicts the Soviet realities of the 1930s, strengthens its position in the most important world cultural heritage, since it still excites foreign readers' interest. Thus, the cult novel was translated not only into European languages, such as German ["Der Meister und Margarita"], Polish ["Mistrz i Małgorzata"], Latvian ["Meistars un Margarita"], English ["Master and Margarita"], Czech ["Mistr a Markéta"], Hungarian ["A Mesterés Margarita"], Spanish ["El maestro y Margarita"], etc., but also into languages of rare language systems, including Mongolian ["master, Margarita hojor"], Chinese ["Dashi Yu magelite"], Japanese ["Kyosyo to Marugarita"], Catalan ["El Mestre I Margarida"], Moldovan ["maestrul şi Margarita»], Macedonian ["Majstro and Margarita"], etc. In addition, it was even translated into Esperanto by Sergei Pokrovsky ("La Majstro kaj Margarita"). As you can see, reading Bulgakov goes far beyond intellectual groups of Russia and is widespread over the world.

Moreover, the large-scale network of the plot prevails in the novel, each of which is filled with its own originality, specificity of characters and details of everyday life. According to the Latvian poet O. Vatsietis, whose translation of "The Master and Margarita" is by far the most significant translation of Bulgakov's works into Latvian, the work on the translation of "The Master and Margarita" is a real school and academy [14]. Indeed, this multifaceted novel goes into three layers: the legendary history of the land of Judah, a satirical image of Moscow 30-ies of the last century, and exaggerated mystical grotesque. Thus, a translator who decided to address himself to such a work, should achieve the effect of reading the translation in the same key by a foreigner reader, as the interpretation of a native reader. Turning to the above mentioned texts of translation, we also must explain such a choice for the experiments. It seemed to us remarkable that the selected translations by M. Glenny (1967) and R. Pevear with L. Volokhonsky (1997) are made 30 years apart. This contributes to the consideration of the subject from a new unexpected angle, namely: it makes it possible to trace the inevitable changes in literary and colloquial

English, as there are a lot of dialogues in Bulgakov's novel, as well as references to all sorts of objects of everyday reality.

Let us briefly note that the two selected translations differ somewhat from each other. On the one hand, being a native speaker, M. Glenny chose a translation strategy which seems to be extremely rich, neat and clear. However, upon a closer view it becomes obvious that the transfer of Russian culture, not to mention the subtleties and spirit of the Soviet era, is difficult for him. The same is not true for the work of R. Pevear and L. Volokhonsky. They not only retain the inherent Bulgakov's meanings of each character, each object, each detail, but also successfully adapt all of the above realities to a reader's world outlook: their translation is accompanied with detailed commentaries that help a foreign reader to penetrate into the environment in which M. Bulgakov intended to immerse the reader.

3.2 NLP Tools and Distributional Semantic Algorithms

The aim of our work is to apply the distributional semantic methods to measure similarity of English translations of fiction text in Russian and extract translation equivalents. The distributional models were deliberately chosen as the tools. Distributional semantics deals with the calculation of the coefficient of semantic similarity between certain language items (words, phrases, sentences). Methods based on the distributive features of language items and key concepts of linear algebra allow us to consider the similarity between language items as the distance between n-dimensional vectors that reflect the distribution of language items. The distance between vectors is trivially calculated by various mathematical formulas (the scalar product, the Euclidean distance, etc.). The most common is the cosine measure, which calculates the cosine of the angle between two vectors as the ratio of their scalar product to the product of each of two vectors lengths:

$$similarity = \cos(\theta) = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} [(B]_i)^2}}$$

The high popularity of the cosine similarity measure is due to the fact that it is appropriate as an evaluation measure of linguistic data, in particular, for processing multi-bit vectors that form a matrix, where each vector is a language item, and each of its dimensions is a set of contexts in which this item appears. The use of cosine measure is especially effective for processing sparse vectors, since such vectors require taking into account only non-zero elements.

For this study, we used a model trained on Google news data corpus [21], which includes approximately about 100 billion words. We believe that this model, which contains 300-dimensional vectors for 3 million words and phrases, is the most optimal for the study with focus on the English language. The model is in open access.

We have chosen Word2Vec neural network toolkit as a set of algorithms allowing to process context vectors, to construct sentence vectors and to perform various operations on vectors. This model was proposed by a group of researchers from Google in 2013 [12], [15]. It appropriately proceeds from the tradition of learning distributed representation for words based on neural networks [16]. This makes it possible for each training sentence to convey the model a number of semantically close sentences. The model learns a distributed representation for each word and the probability function for represented word sequences.

Linguistic data analysis is not limited to working with Word2Vec: it comprises several stages. At the first stage, a corpus should be created and aligned. LF Aligner [22] with a simple and user-friendly interface effectively cope with this task. Sentences aligned in this way clearly demonstrate the most appropriate examples for further sample of data. Selection of contexts for comparison was performed on the principle that we aimed to consider the similarity of sentences that look very different. These differences can be grammatical, lexical or syntactic. Therefore, we extracted the sample in terms of inclusion of the greatest number and variety of translation transformations in the pairs of translated sentences.

Further, on the basis of the obtained sample, we classified the transformations that appear in our sample. Each transformation was weighted depending on its role in the sense shifts. That is, the greater role had a sentence, the greater weight was assigned to it. The transformations and their weights are presented in Table 1.

Table 1. Transformations and their weights.

LEXICAL TRANSFORMATIONS		
Type of transformations	**Content**	**Weight**
SYN	*synonyms*	*1*
SYN PHR	*synonymous expressions*	*1*
CONT SYN	*contextual synonyms*	*2*
ASSOC	*Associative relations, except synonyms and antonyms*	*2-3*
CONV	*conversion*	*2*
DER	*derivation*	*2*
ANT	*antonyms*	*2*
SEMANTIC-SYNTACTIC TRANSFORMATIONS		
Type of transformations	**Content**	**Weight**
STSM	*syntactic transformation with the same meaning*	*1-2*
CCT	*concatenation*	*0,5*
ADD	*adding*	*0,5*
SHF	*shuffling*	*2*
TENSE	*tense changes*	*0,5*
POSR	*part of speech transformation*	*1*

The transformations presented in the Table 1 are based on traditional descriptions of synonymic transformations. The hypothesis of our study is that we are able to compare translations on the basis of the same criteria that underlie the transformations in paraphrasing. Thus, we adapted classifications of transformations proposed by Melchuk [17] and Komissarov [18], [19] and reinforced them with paraphrase classification.

Currently, there are several especially valuable corpora of paraphrases. For the English language, there is the Microsoft Research Paraphrase Corpus [23] created about a decade ago. The given corpus is used for training algorithms for paraphrase recognition. The corpus contains 5,800 pairs of sentences, extracted from news sources on the Internet, and manual tagging which indicates whether each pair of sentences captures a paraphrase relation.

There are two paraphrase corpora for Russian. The Paraphraser.ru [24] corpus is based on news texts. The ParaPlag corpus [25] is developed for testing plagiarism detection methods and contains texts of essays. Both corpora are tagged as regards types of paraphrases and thus reflect acceptable transformations and give an account of possible paraphrase generation techniques.

IV. Experiment on Measuring Similarity of English Translations and Translation Equivalents Extraction

This section is devoted to the description of the experiment on measuring similarity of two English translations of the novel by M.A. Bulgakov "The Master and Margarita". When working on extraction of translation correspondences from the parallel corpus compiled in the Sketch Engine corpus-manager [Sketch Engine] we followed a step-by-step plan described below. Thus, the experiment took place in four stages:

1) measuring similarity of translation pairs by Skip-gram algorithm Word2Vec;
2) analysis of transformations in pairs of sentences in terms of paraphrase types;
3) collocation extraction from a parallel corpus via Sketch Engine.

4.1 Semantic similarity of translation pairs

Word2Vec module from gensim library for Python was used in our experiments. As was mentioned above, the experiment employed a model trained on the Google news corpus.

Preparation of the parallel corpus ORIGINAL – TRANSLATION 1 – TRANSLATION 2 involved in our experiments included its alignment by sentences with the help of LF Aligner toolkit. TXT files were submitted for input. The program gave the output files in TXT, TMX and XLS formats with tabs as the result. An example of aligned sentences is shown in Table 2.

Thereafter we imported the gensim library along with the DocSim module, model Google News-vectors-negative300.bin, stop-words and KeyedVectors model via Jupiter server. A sentence from the translation by M. Glenny was entered as a *SOURCE_DOC* which was an input document and one from the translation by R. Pevear and L. Volokhonsky was entered as a TARGET_DOCS which was an output document. The result is the cosine value of the angle between the compared sentence vectors.

4.2. Analysis of the results based on paraphrase transformation types

We compared translation pairs of sentences in our corpus and tagged them with transformation labels. Not only the ways of paraphrasing were taken into account, but also the length of sentences, direct speech form, translation of certain proper names. We extracted a sample of those translation pairs where the most diverse interpretations were noted. Each pair was evaluated in terms of transformations, the weights being assigned according to our scale, cf. Table 1.

However, during the evaluation process it became obvious that it is incorrect to compare sentences with a significant difference in length by the criterion of the number of transformations. Thus, for example, for sentences of no more than 7 words (without function words), the number of transformations, as a rule, is always less than for sentences of almost 20 words, so that the value of the cosine distance in long sentences also falls. In the final ranking of all the sentences by cosine, there is a serious violation of the regularity put forward in our hypothesis: the greater the cosine of the angle between the vectors, the fewer transformations appear in the sentences. Therefore, we decided to divide the sample into subgroups depending on the number of words. Having done this, we got 12 pairs of 3–7 words; 7 pairs of 8–12 words; 7 pairs of 13–18 words; 3 pairs of 19–25 words; 7 pairs of 30–25 words and within each group we ranked the sentences by increasing the cosine value (Tables 3–7).

Table 2. A Fragment of the aligned corpus

N	Bulgakov	Glenny	Pevear and Volokhonsky
446	«– А–а!	‘Aha!	‘Aha!
447	Вы историк?» – с большим облегчением и уважением спросил Берлиоз.	So you’re a historian?’ asked Berlioz in a tone of considerable relief and respect.	You’re a historian?’ Berlioz asked with great relief and respect.
448	– Я – историк, –подтвердил ученый и добавил ни к селу ни к городу:	‘Yes, I am a historian’, adding with apparently complete inconsequence ,	‘I am a historian,’ the scholar confirmed, and added with no rhyme or reason:
449	– Сегодня вечером на Патриарших прудах будет интересная история!	‘this evening a historic event is going to take place here at Patriarch’s Ponds’.	This evening there will be an interesting story at the Ponds!’
450	И опять крайне удивились и редактор и поэт,	Again the editor and the poet showed signs of utter amazement,	Once again editor and poet were extremely surprised,
451	а профессор поманил обоих к себе и, когда они наклонились к нему,	but the professor beckoned to them and when both had bent their	but the professor beckoned them both to him, and when they leaned towards

	прошептал:	heads towards him he whispered :	him, whispered:
452	– Имейте в виду, что Иисус существовал.	'Jesus did exist, you know.'	‘Bear in mind that Jesus did exist.’
453	– Видите ли, профессор, – принужденно улыбнувшись, отозвался Берлиоз, – мы уважаем ваши большие знания, но	'Look, professor,' said Berlioz, with a forced smile, ' With all respect to you as a scholar we take a	‘You see. Professor,’ Berlioz responded with a forced smile, ‘we respect your great learning, but on this
	сами по этому вопросу придерживаемся другой точки зрения.	different attitude on that point.'	question we hold to a different point of view.’

```
from gensim.models.keyedvectors import KeyedVectors
from DocSim import DocSim

# Using the pre-trained word2vec model trained using Google news corpus of 3 billion running words.
# The model can be downloaded here: https://bit.ly/w2vgdrive (~1.4GB)
# Feel free to use to your own model.
googlenews_model_path = './data/GoogleNews-vectors-negative300.bin'
stopwords_path = "./data/stopwords_en.txt"

model = KeyedVectors.load_word2vec_format(googlenews_model_path, binary=True)
with open(stopwords_path, 'r') as fh:
    stopwords = fh.read().split(",")
ds = DocSim(model,stopwords=stopwords)

source_doc = "Definitely a weird character."
target_docs = ["Ah, what a strange specimen."]
sim_scores = ds.calculate_similarity(source_doc, target_docs)

print(sim_scores)
```

```
[{'score': 0.6605474, 'doc': 'Ah, what a strange specimen.'}]
```

Figure 1. Screenshot of the code calculating a cosine measure [20]

Table 3. Ranked sentences of 3–7 words

Text	**The content of transformations**		**Final weight of transformations**	**COS**
	Lexical	**Semanic-syntactic**		
1) Nothing, except that he was bald and horribly talkative. 2) Nothing except that he was bald and terribly eloquent.	syn (1+1)	-	2	0.917
1) My nerves are in a terrible state. 2) My nerves are really upset, though!	assoc(2)	STSM(1), add(0,5)	3,5	0.761
1) There was silence. Berlioz went pale. 2) Silence fell, and Berlioz paled.	-	STSM(2), CCT(0,5) #concatenation 2->1, add(0,5+0,5)	3,5	0.738
1)Definitely a weird character. 2)Ah, what a strange specimen.	syn(1+1)	STSM(2), add(0,5)	4,5	0.661
1)A lie from beginning to end. 2)A lie from first word to last.	cont syn(2+2)	add(0,5)	4,5	0.650

Table 4. Ranked sentences of 8–12 words

Text	**The content of transformations**		**Final weight of transformations**	**COS**
	Lexical	**Semanic-syntactic**		

1)Rimsky punched himself on the head, spat with fury and jumped back from the window. 2)Rimsky beat himself on the head with his fist, spat, and leaped back from the window.	assoc(2)#punched -beat; syn(1)	-	3	0.840
1)The room smelled of perfume and from somewhere there came the reek of a hot iron. 2)The room smelled of perfume. Besides that, the smell of a red-hot iron was coming from somewhere.	Syn(1+0,5)	CCT(0,5), tense(0,5), SHF(1)	3,5	0.774
1)The mist that came from the Mediterranean sea blotted out the city that Pilate so detested. 2)The darkness that came from the Mediterranean Sea covered the city hated by the procurator.	Assoc(1), syn(1+1)	STSM(1)	4	0.768
1)With an effort Margarita opened it and saw that it contained a greasy yellowish cream. 2)Having mastered herself, Margarita opened it and saw in the box a rich, yellowish cream.	syn phr(1+1), syn(1)	STSM(2)	5	0.729

Table 5. Ranked sentences of 13–18 words

Text	**The content of transformations**		**Final weight of transformations**	**COS**
	Lexical	**Semantic-syntactic**		
1)Suddenly what had delighted him yesterday as proof of his fame and popularity no longer gave the poet any pleasure at all. 2)But the proof of fame and popularity, which yesterday had delighted the poet, this time did not delight him a bit.	syn phr(1)	add(0,5+0,5), SHF(1)	3	0.898
1)Over Moscow it was as if the sky had blossomed: a clear, full moon had risen, still white and not yet golden. 2)The sky over Moscow seemed to lose colour, and the full moon could be seen quite distinctly high above, not yet golden but white.	syn phr (1+1) + ant(2)#blossom -lose colour	add(0,5)	4,5	0.806
1)When he had drunk his hot milk, Ivan lay down again. He was amazed to notice how his mental condition had changed. 2)Having drunk some hot milk, Ivan lay down again and marvelled himself at how changed his thinking was.	syn phr (1), syn(1)	CCT(0,5), tense(1), POSR(1), add(0,5+0,5+0,5)	6	0.747
1)The man was seven feet tall but narrow in the shoulders, incredibly thin and with a face made for decision. 2)A citizen seven feet tall, but narrow in the shoulders, unbelievably thin, and, kindly note, with a jeering physiognomy.	syn(1+1), syn phr(1) + assoc(3) #concretization 'face-physiognomy'	add(0,5+0,5)	7	0.679

Table 6: Ranked sentences of 19–25 words

Text	**The content of transformations**		**Final weight of transformations**	**COS**
	Lexical	**Semantic-syntactic**		
1)A National Library has unearthed some original	conv(1), syn(0,5	STSM(1)	3,5	0.865

manuscripts of the ninth-century necromancer Herbert Aurilachs. I have been asked to decipher them. 2)In a state library here some original manuscripts of the tenth-century necromancer Gerbert of Aurillac 26 have been found. So it is necessary for me to sort them out.	#National-State +1)			
1)Here is my card, my passport and a letter inviting me to come to Moscow for consultations,' said the stranger gravely, giving both writers a piercing stare. 2)Here is my card, my passport, and an invitation to come to Moscow for a consultation,' the stranger said weightily, giving both writers a penetrating glance.	syn phr(1), syn (1+1+1)	-	4	0.797
1)Waking, Margarita did not burst into tears, as she frequently did, because she had woken up with a presentiment that today, at last, something was going to happen. 2)On awakening, Margarita did not weep, as she often did, because she awoke with a presentiment that today something was finally going to happen.	syn phr(1), syn (1+1+1)	tense(0,5)	4,5	0.766

Table 7. Ranked sentences of more than 30 words

Text	**The content of transformations**		**Final weight of transformations**	**COS**
	Lexical	**Semantic-syntactic**		
1) This valuable piece of information had obviously made a powerful impression on the traveller, as he gave a frightened glance at the houses as though afraid of seeing an atheist at every window. 2) The important information apparendy had indeed produced a strong impression on the traveler, because he passed his frightened glance over the buildings, as if afraid of seeing an atheist in every window.	syn (1+1+1+1+1+1+1+1+1)	add(0,5), tense(0,5)	9	0.887
1)He missed his grip and his foot slipped on the cobbles as inexorably as though on ice. As it slid towards the tramlines his other leg gave way and Berlioz was thrown across the track. 2)And right then his hand slipped and slid, one foot, unimpeded, as if on ice, went down the cobbled slope leading to the rails, the other was thrust into the air, and Berlioz was thrown on to the rails.	cont syn phr(2+1+2), cont syn(2), syn(1+1)	CCT(0,5), add(0,5), SHF(1)	11	0.763
1)Leaning back on her comfortable upholstered seat in the trolley-bus, Margarita Nikolayevna rolled along the Arbat, thinking of her own affairs and half-listening to what two men on the seat in front were whispering. 2)Leaning against the comfortable soft back of the trolley-bus seat, Margarita Nikolaevna rode down the Arbat, now thinking her own thoughts, now listening to the whispers of two citizens sitting in front of her.	syn(1+1+0,5+1), assoc(3), cont syn(2)	STSM(1+1), POSR(1), SHF(1)	12,5	0.758

Table 8. Pearson correlation for sentences of 3–7 and 8–12 words

3 – 7 words					**8 – 12 words**				
Weights of the transfor-mations	*COS*	*N*	*Pearson correlation*	*Relevance*	*Weights of the transfor-mations*	*COS*	*N*	*Pearson correlation*	*Relevance*
2 3,5 3,5 4,5 4,5 5 6 6 6,5 7 4 6	0.917 0.761 0.738 0.661 0.650 0.614 0.480 0.479 0.468 0.359 0.299 0.217	12	0,802	0.002	3 3,5 4 5 5,5 5,5 6	0.840 0.774 0.768 0.729 0.658 0.632 0.613	7	0,968	0.000

So, the results satisfy expectations completely: the greater the number of transformations in pairs of sentences, the farther from each other the translation vectors are located, hence the smaller the cosine value.

However, the observed regularities of associations did not satisfy us considerably. Therefore we calculated the Pearson correlation, which is used to determine how proportional the changes of two variables X and Y via the existence of a linear relationship between the two values. For this purpose, we took the weights of the transformations as X and the cosine value as Y.

Table 9: Pearson correlation for sentences of 13–18, 19–25 words and more than 30 words

13-18 words					**19-25 words**				
Weights of the transfor-mations	*COS*	*N*	*Pearson correlation*	*Relevance*	*Weights of the transfor-mations*	*COS*	*N*	*Pearson correlation*	*Relevance*
3 4,5 6 7 7 8,5 9	0.898 0.806 0.747 0.679 0.678 0.597 0.507	7	0,968	0.000	3,5 4 4,5	0.865 0.797 0.766	3	0,973	0.004
>30 words									
9 11 12,5…	0.887, 0.763, 0.758 …	**7**	0,978	0.003					

Table 10: A Fragment of collocation lists

Bulgakov	**Glenny**	**Pevear and Volokhonsky**	**Bulgakov**	**Glenny**	**Pevear and Volokhonsky**
сиплым голосом	nasal voice	nasal voice	бросились к	went straight to	dashed to
белом плаще	white cloak	white cloak	подсолнечное масло	sunflower-seed oil	sunflower oil
верхней террасе	upper terrace	upper terrace	работал усердно	worked hard	worked assiduously
тайной службы	secret service	secret service	раздвоение ивана	the two ivans	ivan splits in two

As we can see in Tables 8-9, the results turned up to be satisfactory. According to the authors of the SPSS, we should rely on the relevance of the correlation coefficient when evaluating results. The relevance index is lower than 0.05, hence the correlation is relevant. The coefficient of Pearson correlation itself also highlights a strong association between variables *COS* and *Weight of transformations*.

4.3. Collocation extraction

The third stage of our experiment required processing of the original text of the novel and its translations. Our task came down to detecting, by trial and error, the optimal way to identify collocations from all three sources, and then, ordering them by decreasing the association strength to observe whether there are any translation equivalents among them.

Collocation extraction was performed by means of Sketch Engine [Sketch Engine] corpus-manager. It allows to extract a list of collocations from a corpus by selected statistical parameters and using stop words. Thus, the program selected by default 1000 n-grams for each of our corpora. After loading a stop-word list, the number of n-grams decreased to 159, 143 and 185 for Bulgakov's text in Russian, Glenny's translation and Pevear's translation respectively. Examples of obtained collocations with frequency are presented in Table 10.

Having ordered the obtained collocations according to the degree of relevance and frequency in reference to the corpora, we obtained satisfactory results. First of all, the n-grams from all three texts are identified correctly with adduction of their frequency. There were only minor inaccuracies, namely the variants "n't understand, n't help" etc. and "ye gods". Especially successful proper names as "Vasily Stepanovich – Vassily Stepanovich" or "variety theatre" are selected in the same way.

Any intersections within the three texts are so desirable for us, as they can be used as translational equivalents for lemmas and constructions in bilingual dictionaries. Thus, 43% of the collocations are the same in all three documents; 25% of the collocations appear in only two of the three documents. The coincidences from the *original-translation1 text* and the *original-translation2 text* can be considered as interchangeable and used as new translations for a particular language unit.

V Conclusion

In the given paper we described the experimental study involving application of distributional semantic models in comparison of two English translations. The experiment was performed with the text of M.A. Bulgakov's novel "The Master and Margarita" and two of its translations into English.

In general, the results of the study answer the expectations:

1) we proved that the procedure for comparing translations can be implemented via linear and vector metrics used in the measuring similarity of texts;
2) we confirmed the hypothesis that the translated texts themselves can be considered as paraphrases obtained as a result of some transformations;
3) the results provided by language models built by Word2Vec algorithms received very close quality estimates.

Thus, our study confirms the effectiveness of Word2Vec neural network models for measuring similarity of two translations. The identification of collocations from text corpora as a variation of the extraction of translation equivalents also allowed to obtain a high rate of coincidences.

Acknowledgment

I would like to thank PhD, Associate Professor O.A. Mitrofanova (Saint Petersburg State University) for fruitful discussion of this research project and for comments that greatly improved this paper.

References

[1] K. Church, P. Hanks, "Word Association Norms, Mutual Information, and Lexicography". *Computational Linguistics,* vol. 16, issue 1, 1990, pp. 22–29.

[2] F. Smadja, "Retrieving collocations from text: Xtract". *Computational Linguistics - Special issue on using large corpora,* vol. 19, issue 1, 1993, pp. 143–177.

[3] D. Lin, "Using collocation statistics in information extraction". *Proceedings of the Seventh Message Understanding Conference (MUC-7), 1998.*

[4] St. Evert, "Corpora and Collocations". *Corpus Linguistics. An International Handbook / A. Lüdeling, M. Kytö (eds.),* 2008, article 58, pp. 1212–1248.

[5] V.Seretan, Syntax-Based Collocation Extraction. Text, Speech and Language Technology series, vol. 44, 2011.

[6] L. Wanner, B. Bohnet, M. Giereth, "Making sense of collocations". *Computer, Speech and Language,* vol. 20, issue 4, 2006, pp. 609–624.

[7] J. Kupiec, "An algorithm for finding noun phrase correspondences in bilingual corpora". *Proceedings of the 31st annual meeting on association for computational linguistics (ACL 1993),* 1993, pp. 17–22.

[8] M. Haruno, S. Ikehara, T. Yamazaki, "Learning bilingual collocations by word-level sorting". *Proceedings of the 16th Conference on Computational Linguistics,* vol. 1, 1996, pp. 525–530.

[9] Ch.-Ch. Wu, J.S. Chang, "Bilingual collocation extraction based on syntactic and statistical analyses". *Proceedings of the 15th Conference on Computational Linguistics and Speech Processing. Association for Computational Linguistics and Chinese Language Processing,* 2003, pp. 1–20.

[10] P. Fung, “A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora”. *Proceedings of the 3rd Conference of the Association for Machine Translation in the America. Machine Translation and the Information Soup (AMTA 1998),* 1998, pp. 1–17.

[11] G. Bukia, E. Protopopova, O. Mitrofanova, “A corpus-driven estimation of association strength in lexical constructions”. *Proceedings of the AINL-ISMW FRUCT, FRUCT Oy, Finland,* 2015, pp. 147–152.

[12] T. Mikolov, K. Chen, G. Corrado, J. Dean, “Efficient estimation of word representations in vector space”. *Workshop Proceedings of the International Conference on Learning Representations (ICLR),* 2013.

[13] Yu. Morozova, E. Kozerenko, M. Sharnin, “Method for extracting single-word translation correspondences from parallel texts using distributional semantics models”. *Systems and Means of Informatics,* vol. 24., issue 2, 2014, pp. 131–142. (In Rus.) = Yu. Morozova, E. Kozerenko, M. Sharnin, “Metodika izvlechenija poslovnyh perevodnyh sootvetstvij iz parallelnyh tekstov s primenenijem modelej distributivnoj semantiki”. *Sistemy i sredstva informatiki,* tom 24, vyp. 2, 2014. Pp. 131–142.

[14] O. Vācietis, *Ieejam Bulgakova galaktikā. Jaunās grāmatas,* № 11, 1979. (In Lat.) = O. Vācietis, *Vhodim v galaktiku Bulgakova.* Novyje knigy, № 11, 1979.

[15] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, “Distributed representations of words and phrases and their compositionality”. *NIPS’13 Proceedings of the 26th International Conference of Neural Information Processing Systems, 2013.*

[16] Y. Bengio, “A Neural Probabilistic Language Model”. *Journal of Machine Learning Research 3*, 2003, pp. 1137–1155.

[17] Melchuk, *The experience of the theory of linguistic models “Meaning <=> Text”.* M., 1999. (In Rus.) = I. Melchuk, *Opyt teorii lingvisticheskih modeley “Smysl<=> Tekst”.* M., 1999.

[18] V. Komissarov, *Theory of translation*. M., 1990. (In Rus.) = V. Komissarov, *Teorija perevoda*. M., 1990.

[19] V. Komissarov, *Modern translation science.* M., 2004. (In Rus.) = V. Komissarov, *Sovremennoje perevodovedenije. M., 2004*.

[20] DocSim: the code for the cosine measure. URL: https://github.com/v1shwa/document-similarity

[21] Google News Corpus. URL: https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit.

[22] LF Aligner. URL: https://sourceforge.net/projects/aligner/

[23] Microsoft Research Paraphrase Corpus. URL: https://www.microsoft.com/enus/download/details.aspx?id=52398&from=http%3A%2F%2Fresearch.microsoft.com%2Fen-us%2Fdownloads%2F607d14d9-20cd-47e3-85bc-a2f65cd28042%2Fdefault.aspx

[24] ParaPhraser. URL: http://paraphraser.ru/

[25] ParaPlag. URL: https://plagevalrus.github.io/content/corpora/paraplag.html