Designing a Russian Idiom-Annotated Corpus

Katsiaryna Aharodnik, Anna Feldman, Jing Peng
CUNY, The Graduate Center, Montclair State University
New York, New York, Montclair, New Jersey, United States
kaharodnik@gradcenter.cuny, {feldmana,pengj}@montclair.edu

Abstract

This paper describes the development of an idiom-annotated corpus of Russian. The corpus is compiled from freely available resources online and contains texts of different genres. The idiom extraction, annotation procedure, and a pilot experiment using the new corpus are outlined in the paper. Considering the scarcity of publicly available Russian annotated corpora, the corpus is a much-needed resource that can be utilized for literary and linguistic studies, pedagogy as well as for various Natural Language Processing tasks.

Keywords: idioms, annotation, corpus, Russian.

1. Introduction

In recent years, there has been a growing interest in exploring the questions of automatic processing of semantic relationships and specifically those that are not trivial to define and disambiguate. Among these questions is the problem of automatic identification of figurative language within a large body of text. Largely, the problem lies in the ambiguous nature of idiomatic expressions and identifying the cues for idiom recognition. Some expressions can be interpreted either literally or idiomatically depending on the context in which they occur. Several approaches have been explored in finding a better solution to this problem (e.g., Fazly et al., 2009; Cook et al., 2007; Katz and Giesbrecht, 2006; Sporleder & Li, 2009; Li & Sporleder, 2010; Pradhan et al., 2017; Peng & Feldman, 2016(a, b); Peng et al., 2015; Peng et al., 2014, among others). Unfortunately, the corpora that could be used for training idiom classifiers are scarce, especially if one turns to languages other than English.

In this paper, we describe an idiom-annotated corpus for Russian. This corpus is a valuable language resource which can be used for various Natural Language Processing (NLP) tasks, such as automatic idiom recognition. Also, it can be utilized as a pedagogical tool for teaching the intricacies of the Russian language or as a corpus for linguistic investigations. Our corpus is available for research purposes https://github.com/kaharodnik/Ru_idioms. A pilot experiment using the idiom-annotated corpus is also described in the paper.

2. Motivation

Idioms lack a clear observable relation between the linguistic meaning and interpretation. Moreover, expressions can be ambiguous between idiomatic and literal interpretation depending on the context in which they occur (e.g., *sales hit the roof* vs. *the roof of the car*). Fazly et al.'s (2009) analysis of 60 idioms from the British National Corpus (BNC) has demonstrated that close to half of such expressions have a clear literal meaning; and of those with a literal meaning, on average around 40% of their usages are literal. Therefore, idioms present great challenges for many NLP applications, such as machine translation.

There has been substantial computational research on idioms, with an emphasis on English. Previous approaches to idiom detection can be classified into two groups: 1) type-based extraction, i.e., detecting idioms at the type level; 2) token-based detection, i.e., detecting idioms in context. Type-based extraction relies on the idea that idiomatic expressions exhibit certain linguistic properties such as non-compositionality that can distinguish them from literal expressions (Sag et al., 2002; Fazly et al., 2009). While many idioms can be characterized by these properties, a number of idioms fall on the continuum from being compositional to being partly unanalyzable to completely non-compositional (Cook et al., 2007). Katz and Giesbrecht (2006), Birke and Sarkar (2006), Fazly et al. (2009), Sporleder and Li (2009), Li and Sporleder (2010), among others, emphasize that type-based approaches do not work on expressions that can be interpreted either idiomatically or literally depending on the context, and thus an approach that considers tokens in context is more appropriate for idiom recognition. Different token-based approaches have been proposed for more efficient ways of idiom identification. Some of them use topic-based representation (Peng et al. 2014); others utilize word embeddings (Peng et al., 2015, 2016; Pradhan et al., 2017). The above approaches rely on corpora annotated for both literal and idiomatic interpretations of expressions. The corpus proposed in this paper, besides its more general purpose, satisfies this requirement and thus is an important contribution to the community of researchers working on idiom detection in general and on Russian idioms in particular.

3. Corpus Description

Following the rationale for token-based approach, each corpus entry contains a target expression itself (idiomatic or literal) and two paragraphs of context. Thus, each entry is divided into three paragraphs: one paragraph preceding the paragraph with a target expression and the other following the paragraph with a target expression. Each target expression can be identified as both, idiomatic or literal, depending on the context. Each file of the corpus contains one entry. The examples of two corpus entries below show one-paragraph entries for literal (L) and idiomatic (I) interpretations of a target expression ***на чемоданах*** (*na čemodanah*) - *on suitcases.* Example 1, Literal:

*Народ табором расположился **на чемоданах** и баулах, расслабленно сидел, опустив руки, а кто-то доставал походную снедь, по палубе расползались ароматы жареных кур и копченой рыбы. У судна стали собираться крикливые чайки.*

In the above example, the target expression ***на чемоданах*** *(na čemodanah) on suitcases* is located in the second paragraph of the corpus entry. It can be interpreted literary *to sit on suitcases*. In the corpus entry below, the same target expression is interpreted idiomatically *to be packed and waiting, to be unsettled.* Generally, this idiom is similar to the English idiom *to live out of a suitcase.* Example 2, Idiomatic:

*Шло время, но разрешения из ОВИРа не приходило. Афганская кампания ввода ограниченного контингента войск смешала все карты. Запах холодной войны проникал в самые отдаленные сферы жизни и прежде всего в государственную политику по так называемому тогда воссоединению семей. Единственная законная возможность уехать из страны Советов все более переходила в область мифов. Казалось, что выезд закрыт навсегда. Ждать всегда противно, а ждать разрешения на выезд противно вдвойне. Сколько времени можно жить **на чемоданах**? Год, два, десять? Тем, кто работал сторожами и лифтерами, было вообще грустно: ни работы нормальной, ни перспектив.*

These examples demonstrate that an entry provides substantial context for each target expression in the corpus. The preceeding paragraph and the one following it are omitted in the examples.

To make the corpus balanced across written registers, it was compiled from texts of different genres: fiction and non-fiction, Wikipedia style text. The fiction sub-corpus was also split into two parts: Classical Russian Literature and Modern Russian Literature. The texts for this part were extracted from freely available online Russian library, Moshkov's library (http://lib.ru/). Classical literature texts were taken from Классика(Classical)/Проза(Prose). This part of corpus consists of Russian prose of late nineteenth-early twentieth century. Similarly, Modern literature sub-corpus consists of prose from Современная (Modern)/Проза (Prose) part of the library. In Modern Prose, the texts are written by a variety of Russian authors dating back to the second half of the twentieth and twenty-first centuries. The Wikipedia sub-corpus (Ru Wiki) was created from Russian Wikipedia freely available at http://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/. In the corpus, the files were saved in folders according to genres, making it possible for researchers to conduct comparative analyses. Each text for Classical and Modern literature sub-corpora was saved in a separate file. The Ru Wiki sub-corpus was analyzed as a single XML file. Table 1 describes the total number of tokens used for idioms extraction for each part of the corpus.

Once the Russian corpus was compiled, the list of target expressions (idioms) of interest was created (see Section 4).

Corpus	# tokens
Classical Prose	111,725,751
Modern Prose	46,996,232
Ru Wiki	486,474,989

Table 1: Description of Sub-Corpora

4. Target Expressions

For the list of idioms, a Russian-English dictionary of idioms was used as a primary source (Lubensky, 2013). Initially, 150 idioms (target expressions) were included in the list. The rationale for choosing a certain target expression was that each expression could be interpreted as either idiomatic or literal depending on the context. Some idioms were not found in the source files and were excluded from the list. The final list consisted of 100 target expressions. This final list was used for compiling the actual annotated corpus.

The list of idioms included only multiword expressions (MWE). Each target expression consisted of more than one-word token, with their length ranging from two-word tokens, e.g., *длинный язык- long tongue*, to four-word tokens as in *с пеной у рта – with frothing at the mouth*. Syntactically, target expressions were not limited to a single structure. They could be separated into three groups: Noun Phrases (NP), Prepositional Phrases (PP), and Verb Phrases (VP) types of constructions. The PP type included Preposition + Noun, e.g., *без головы* (*without the head*), Preposition + Adjective/Attributive Pronoun + Noun, e.g., *на свою голову (on one's head)*, the NPs included Adjective/Possessive Pronoun + Noun e.g., *второй дом* (*second home*), and VP type included Verb + Preposition + Noun, e.g., *плыть по течению* (*to go with the flow*), and Verb + Noun, e.g., *поставить точку* (*to put a stop*). Table 2 provides a list of syntactic constructions with their counts. The list included idioms in their dictionary form, but each idiomatic expression was extracted from the compiled corpora in any form it appeared in files (conjugated forms for verbs or declined forms for adjectives and nouns).

4.1 Extracting Target Expressions

A target token is defined as a multiword expression that can be identified as either idiomatic or literal within the text. Each target expression was extracted with one preceding and one following paragraph from a source text file. Thus, one entry is defined as a three-paragraph text in one file.

Syntactic Construction	Russian	English	Count
Adj (Poss Pron) + Noun	Черный ворон	Black raven	33
Prep+Noun	Без головы	Without the head	82
Prep+Adj+Noun	На мою голову	On my head	78
Verb+(Prep)+Noun	Вцепиться в глотку	To grab one's throat	50
Adv + Verb	Жирно будет	Too greasy (too much)	9
Noun + Short Adj	Концерт Окончен	The concert is over	4
Prep+Noun+Verb	Куда ветер дует	Where the wind blows	7

Table 2: Syntactic Constructions of Idiomatic Expressions

Each target expression was extracted following the steps below:

1. Convert the online text file to html format. This was done to preserve the html tags and use the tags for paragraph extraction.
2. Save each file as a plain text document with preserved html tags.
3. Extract each target expression (token) from each html document in a three-paragraph format, with the second paragraph containing a target expression.
4. Save each three-paragraph entry in a separate text file.

Overall, 100 tokens/target expressions were used to create the idiom-annotated corpus. The number of files in each sub-corpus varied depending on the amount of the idiomatic/literal expressions found in the sub-corpora.

4.2 Annotation

Once the expressions were extracted, each file was annotated manually by two Russian native speakers with overall high inter-annotator agreement (Kappa 0.81). Each target expression was assigned a tag Idiomatic (I) or Literal (L). Once the annotator made a decision about the tag, the three paragraph entries were saved in a text file format. In some cases, the resulting files did not have a required amount of paragraphs and were marked as a no paragraph label *_np* within a file name, e.g., *na_moyu_golovu_I_3_np.txt*. This could have happened for several reasons. Sometimes, preceding or following paragraphs could have been contaminated with tags without a sufficient amount of actual text. In these cases, the files were cleaned to include only intelligible text. In other cases, the target expressions were found in the first or last paragraph of a source file, hence they were missing the required amount of context. However, these files were not excluded from the corpus, since they can be still used for the analyses. The list of 10 most frequent target expressions extracted for the corpus is provided in Table 3. Table 3 also includes the counts of idiomatic and literal interpretations for each idiom. For each entry, an XML file was created with a label for an idiomatic expression within a file.

As the result, the idiom-annotated Russian corpus contained the three sub-corpora of files in plain text and XML formats with each target expression, three paragraph entries per file. The annotators' labels are assigned within XML files and are reflected in the folder names for plain text files. README files are also provided for each sub-corpora. Each README file lists the file directory for an idiomatic expression (File folder/File Name), the corresponding target expression in Russian, its translation in English, and the number of tokens (words and punctuation) prior to the first token of the idiomatic expression. The total counts for literal and idiomatic expressions extracted per sub-corpora are listed in Table 4.

#	Target	Gloss	Interpretation	I	L
1	s bleskom	with flying colors	brilliantly	246	78
2	na svoju golovu	on your own head	pain in the neck	185	58
3	na vysote	at the height	rise to the occasion	294	438
4	smotret' v glaza	look into the eyes	face (challenges)	48	83
5	čerez golovu	over the head	go over someone's head	100	316
6	na nožax	with the knives	to be at daggers drawn	53	43
7	po barabanu	on the drums	couldn't care less	86	25
8	vtoroj dom	second home	second home	14	40
9	vyše sebja	above oneself	beyond the possible	57	22
10	dlinnyj jazyk	long tongue	chatterbox	37	29

Table 3: Ten most frequent target expressions.

Sub-Corpus	# Literal Expressions	# Idiomatic Expressions	#Total files
Classical Literature	2,100	1,231	3,331
Modern Literature	612	803	1,415
Russian Wiki	315	386	701

Table 4: Literal and Idiomatic Total Counts per Sub-Corpora.

5. Idiom Detection Experiment

Below we report the results of a pilot idiom detection experiment for which we used the idiom-annotated corpus described in this paper. For this pilot experiment, we follow the hypotheses and the methodology described in Peng et al. (2018). The automatic idiom detection approach is based on two hypotheses: (1) words in a given text segment that are representatives of the local context are likely to associate strongly with a literal expression in the segment, in terms of projection of word vectors onto the vector representing the literal expression; (2) the context word distribution for a literal expression in word vector space will be different from the distribution for an idiomatic one (similarly to Firth, 1957; Katz and Giesbrecht, 2006).

5.1 Projection based on Local Context Representation

To address the first hypothesis, we propose to exploit recent advances in vector space representation to capture the difference between local contexts (Mikolov et al., 2013a; Mikolov et al., 2013b).

A word can be represented by a vector of fixed dimensionality q that best predicts its surrounding words in a sentence or a document (Mikolov et al., 2013a; Mikolov et al., 2013b). Given such a vector representation, our first proposal is the following. Let v and n be the vectors corresponding to the verb and noun in a target verb-noun construction, as in *blow whistle*, where $v \in \Re^q$ represents *blow* and $n \in \Re^q$ represents *whistle*. Let

$$\sigma_{vn} = v+n \in \Re^q.$$

Thus, σ_{vn} is the word vector that represents the composition of verb v and noun n, and in our example, the composition of *blow* and *whistle*. As indicated in (Mikolov et al., 2013b), word vectors obtained from deep learning neural net models exhibit linguistic regularities, such as additive compositionality. Therefore, σ_{vn} is justified to predict surrounding words of the composition of, say, *blow* and *whistle* in a literal context. Our hypothesis is that on average, the projection of v onto $\sigma_{blowwhistle}$, (i.e., $v{\cdot}\sigma_{blowwhistle}$, assuming that $\sigma_{blowwhistle}$ has unit length), where vs are context words in a literal usage, should be greater than $v\ {\cdot}\sigma_{blowwhistle}$, where vs are context words in an idiomatic usage.

For a given vocabulary of m words, represented by matrix

$$V = [v_1,v_2,\cdots,v_m] \in \Re^{q\times m},$$

We calculate the projection of each word v_i in the vocabulary onto σ_{vn}

$$P = V^t\sigma_{vn} \qquad (1)$$

where $P \in \Re^m$, and t represents transpose. Here we assume that σ_{vn} is normalized to have unit length. Thus, $P_i = v^t_i\sigma_{vn}$ indicates how strongly word vector v_i is associated with σ_{vn}. This projection forms the basis for our proposed technique.

Let $D = \{d_1,d_2,\cdots,d_l\}$ be a set of l text segments (local contexts), each containing a target VNC (i.e., σ_{vn}). Instead of generating a term by document matrix, where each term is *tfidf* (product of term frequency and inverse document frequency), we compute a term by document matrix

$M_D \in \Re^{m\times l}$, where each term in the matrix is

$$p{\cdot}id\,f. \qquad (2)$$

That is, the product of the projection of a word onto a target VNC and inverse document frequency. That is, the term frequency (tf) of a word is replaced by the projection of the word onto σ_{vn} (1). Note that if segment d_j does not contain word v_i, $M_D(i,j)= 0$, which is similar to *tf-idf* estimation. The motivation is that topical words are more likely to be well predicted by a literal VNC than by an idiomatic one. The assumption is that a word vector is learned in such a way that it best predicts its surrounding words in a sentence or a document (Mikolov et al., 2013a; Mikolov et al., 2013b). As a result, the words associated with a literal target will have larger projection onto a target σ_{vn}. On the other hand, the projections of words associated with an idiomatic target VNC onto σ_{vn} should have a smaller value.

We also propose a variant of $p{\cdot}id\,f$ representation. In this representation, each term is a product of p and typical *tf-idf*. That is,

$$p{\cdot}t\,f\,{\cdot}id\,f. \qquad (3)$$

5.2 Local Context Distributions

Our second hypothesis states that words in a local context of a literal expression will have a different distribution from those in the context of an idiomatic one. We propose to capture local context distributions in terms of scatter matrices in a space spanned by word vectors (Mikolov et al., 2013a; Mikolov et al., 2013b).

Let $d = (w_1,w_2\cdots,w_k) \in \Re^{q\times k}$

be a segment (document) of k words, where $w_i \in \Re^q$ are represented by a vectors (Mikolov et al., 2013a; Mikolov et al., 2013b). Assuming w_is have been centered, we compute the scatter matrix

$$\Sigma = d^t d, \qquad (4)$$

where Σ represents the local context distribution for a given target VNC.

Given two distributions represented by two scatter matrices Σ_1 and Σ_2, a number of measures can be used to compute the distance between Σ_1 and Σ_2, such as Choernoff and Bhattacharyya distances (Fukunaga, 1990). Both measures require the knowledge of matrix determinant. We propose to measure the difference between Σ_1 and Σ_2 using matrix norms. We have experimented with the Frobenius norm and the spectral norm. The Frobenius norm evaluates the difference between Σ_1 and Σ_2 when they act on a standard basis. The spectral norm, on the other hand, evaluates the difference when they act on the direction of maximal variance over the whole space.

5.3 Methods

We carried out an empirical study evaluating the performance of the proposed techniques. The following methods are evaluated:

1. $p{\cdot}id\,f$: compute term by document matrix from training data with proposed $p{\cdot}id\,f$ weighting (2).
   $p \cdot t\,f \cdot id\,f$: compute term by document matrix from training data with proposed p*tf-idf weighting (3).
2. $CoVAR_{Fro}$: proposed technique (4) described in Section 2.2, the distance between two matrices is computed using Frobenius norm.
3. $CoVAR_{Sp}$: proposed technique similar to $CoVAR_{Fro}$. However, the distance between two matrices is determined using the spectral norm.

For methods 3 and 4, we compute the literal and idiomatic scatter matrices from training data (4). For a test example, compute a scatter matrix according to (4), and calculate the distance between the test scatter matrix and training scatter matrices using the Frobenius norm for method 3, and the spectral norm for method 4.

5.4 Results

The results of the experiment suggest that for Russian our algorithm performs similarly to English, even considering the fact that Russian is a more morphologically complex language and has a relatively free word order. Specifically, the results demonstrate that one of our proposed methods - $CoVAR_{Fro}$ performs with highest average accuracy for precision and recall measures. The results are described in Table 5.

6. Corpus Importance

In this paper, we described the development of a Russian-language corpus annotated for idioms. This corpus is pivotal for a variety of NLP tasks such as idiom detection, as well as a useful resource for various linguistic analyses and pedagogical applications. The corpus contains only those expressions whose idiomatic or literal interpretation depends on context. The format of the corpus allows the user to easily search for idioms in context. In addition, unlike previous corpora annotated for idioms (e.g., Cook et al., 2008), this corpus contains expressions of various syntactic types.

Method	na svoju golovu get into trouble	na vysote to be at one's best	smotret' v glaza to face (a challenge)	Ave
		Precision		
$p{\cdot}id\,f$	0.75	0.49	0.40	0.55
$p{\cdot}t\,f{\cdot}id\,f$	0.80	0.50	0.50	0.60
$CoVAR_{Fro}$	0.80	0.71	0.49	**0.67**
$CoVAR_{sp}$	0.78	0.64	0.54	0.65
		Recall		
$p{\cdot}id\,f$	0.73	0.83	0.40	0.65
$p{\cdot}t\,f{\cdot}id\,f$	0.76	0.81	0.42	0.66
$CoVAR_{Fro}$	0.88	0.81	0.50	**0.73**
$CoVAR_{sp}$	0.76	0.76	0.50	0.67
		Accuracy		
$p{\cdot}id\,f$	0.63	0.64	0.57	0.61
$p{\cdot}t\,f{\cdot}id\,f$	0.68	0.66	0.67	0.67
$CoVAR_{Fro}$	0.76	0.82	0.65	**0.74**
$CoVAR_{sp}$	0.68	0.77	0.68	0.71

Table 5: Average performance of competing methods on Russian idioms.

More generally, the described corpus facilitates research in the Russian language. Since the corpus contains sections from different time periods and genres, it is possible to investigate the usage of idioms in fiction vs. non-fiction or explore how figurative language changes over time. The variety of grammatical constructions provides insights into the syntactic nature of Russian idioms, especially those that can be productively used in either idiomatic or literal sense.

In this paper, we also reported the results of a pilot experiment using the corpus. The experiment demonstrates the feasibility of using the corpus for automated idiom identification approaches. We are planning to expand the size of the corpus in the future, by extracting more types of target expressions and adding other genres.

7. Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-16-1-0261 and a National Science Foundation grant IIS-1319846.

8. References

Birke, J. and Sarkar, A. (2006). A Clustering Approach to the Nearly Unsupervised Recognition of Nonliteral Language. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06), pages 329–226, Trento, Italy.

Cook, P., Fazly, A., Stevenson, S. (2008). The VNC-tokens dataset. Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008*)*.

Fazly, A., Cook, P., Stevenson, S. (2009). Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics 35* (2009), 61-103

Katz, G., Giesbrecht, E. (2006). Automatic Identification of Non-compositional Multiword Expressions using Latent Semantic Analysis. Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties. (2006) 12-19

Li, L., Sporleder, C. (2010). Using Gaussian Mixture Models to Detect Figurative Language in Context. In: *Proceedings of NAACL/HLT 2010*.

Lubensky, S. (2013). Russian-English Dictionary of Idioms. Yale University Press.

Peng, J., Aharodnik K., and Feldman A. (2018). A Distributional Semantics Model for Idiom Detection: The Case of English and Russian. In the Proceedings of the 10th International Conference on Agents and Artifificial Intelligence (Special Session on Natural Language Processing in Artificial Intelligence - NLPinAI 2018).

Peng, J. and Feldman, A. (2016). Experiments in Idiom Recognition. *Proceedings of the 26th International Conference on Computational Linguistics (COLING). Osaka, Japan*.

Peng, J. and Feldman, A. (2016). In God We Trust. All Others Must Bring Data. — W. Edwards Deming — Using word embeddings to recognize idioms. Proceedings of the 3rd Annual International Symposium on Information Management and Big Data — SIMBig, Cusco, Peru.

Peng, J., Feldman, A., and Jazmati, H. (2015). Classifying Idiomatic and Literal Expressions Using Vector Space Representations. Proceedings of the Recent Advances in Natural Language Processing (RANLP) conference 2015, Hissar, Bulgaria, September 2015

Peng, J., Feldman, A., and Vylomova, E. (2014). Classifying Idiomatic and Literal Expressions Using Topic Models and Intensity of Emotions. In Proceedings of the 2014 Empirical Methods for Natural Language Processing Conference (EMNLP).

Pradhan, M., Peng, J., Feldman, A., and Wright, B. (2017)**.** Idioms: Humans or machines, it's all about context. *In Proceedings of 18th International Conference on Computational Linguistics and Intelligent Text Processing. Budapest. Hungary.*

Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2002). Multiword expressions: A Pain in the Neck for NLP. Proceedings of the 3rd International Conference on Intelligence Text Processing and Computational Linguistics (CICLing 2002), Mexico City, Mexico (2002) 1–15

Sporleder, C., Li, L. (2009). Unsupervised Recognition of Literal and Non-literal Use of Idiomatic Expressions. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09), Morristown, NJ, USA, Association for Computational Linguistics (2009) 754-762