National Research University Higher School of Economics

as a manuscript

Yulia Badryzlova

AUTOMATED METAPHOR IDENTIFICATION IN RUSSIAN TEXTS

Dissertation Summary for the academic degree
of Doctor of Philosophy in Philology and Linguistics HSE

Academic Supervisor:
Olga Lyashevskaya, PhD

Moscow 2019

OVERVIEW OF THE THESIS

Metaphor occupies a prominent place in contemporary linguistic theory: it is recognized to be one of the most powerful cognitive tools with which humans conceptualize (Lakoff & Johnson, 1980a). Evidence from psycholinguistic research demonstrates that metaphor guides reasoning and decision-making in societal, economic, educational, health-related, and environmental issues (Hendricks & Boroditsky, 2016; Thibodeau & Boroditsky, 2011).

Metaphor is truly ubiquitous in everyday discourse; metaphor's pervasiveness is estimated invariably high: on the average, 0.3 metaphor occurs in a sentence in a multi-domain corpus (Shutova & Teufel, 2010); in genre-specific corpora, the frequency of metaphor ranges within 5-18% of the total number of words (Steen et al., 2010).

Not surprisingly, metaphor identification and interpretation pose a serious challenge to a wide range of real-world NLP applications, such as information retrieval, machine translation, question answering, information extraction, opinion mining, and others. The latest advances in corpus linguistics and machine learning sparked a large-scale wave of computational metaphor projects. A series of Workshops on Metaphor in NLP was held for several successive years as a part of the NAACL-HLT conference (Klebanov, Shutova, & Lichtenstein, 2014, 2016; Shutova, Klebanov, & Lichtenstein, 2015; Shutova, Klebanov, Tetreault, & Kozareva, 2013). The first competition of NLP systems in a shared metaphor detection task was held in 2018 (Leong, Klebanov, & Shutova, 2018) where systems were evaluated on the dataset of the VU Amsterdam Metaphor Corpus, VUAMC (Steen et al., 2010).

Metaphor corpora and metaphor identification systems may be primarily divided into the two major groups – those that operate within the theoretic paradigm of conceptual metaphor, on the one hand, and those that do not make any *a priori* assumptions about the underlying conceptual mechanisms of metaphor and focus on linguistic metaphor. Conceptual metaphor is understood as a mapping between two distinct conceptual domains, the Source and the Target. The first approach is "top down", that is, the researcher starts out either from a set of predefined conceptual metaphors or from particular source or target domains. The second approach is "bottom up" – no specific conceptual metaphor is presumed: it aims to identify the linguistic metaphors as surface expressions of possible underlying cross-domain mappings.

The central goal of this thesis is to provide in-depth linguistic analysis of context features which can be utilized in order to automatically differentiate utterances which contain linguistic metaphor from non-metaphoric ones. We address this goal by running several machine learning experiments for metaphor identification in Russian and by evaluating the importance of each of the proposed features; the latter evaluation is also performed using machine learning algorithms. It should be emphasized that the present work does not aim to engineer an algorithm which would maximize the performance on the metaphor identification task; rather, we intend to suggest feature extraction methods and to assess the efficiency of the extracted features.

The main goal of the thesis is accomplished via the following series of tasks:

- to develop a customized scheme for annotation of linguistic metaphor at the sentence level;
- to collect and annotate a corpus of contexts containing linguistic metaphor, as well as non-metaphoric ones;
- to evaluate the quality of metaphor annotation;
- to suggest methods of feature engineering for identification of linguistic metaphor at the sentence level;
- to implement machine learning experiments for linguistic metaphor identification using models based on the suggested features and their combination;
- to evaluate the performance of the models and their generalizability for experiments on new datasets;
- to provide an in-depth linguistic analysis of the contextual factors that promote the success or failure of features.

The types of features to be explored in this research are:

1) Semantic similarity;
2) Lexical co-occurrence;
3) Morphosyntactic co-occurrence;
4) Concreteness indexes;
5) Occurrences of flag words (lexical signals of metaphoricity) and quotation marks.

The following **methods and algorithms** are used in the present research:

- a customized version of the MIPVU procedure for annotation of linguistic metaphor (Steen et al., 2010);
- distributional semantic models (Baroni et al. 2013, Kutuzov & Kuzmenko, 2016);
- ΔP metric, a statistical measure of association (Levshina, 2015);

– Support Vector Machine algorithm;
– Random Forest algorithm;
– Logistic Regression algorithm;
– K-means clustering algorithm;
– Boruta algorithm (Kursa, Jankowski, & Rudnicki, 2010);

Relevance the thesis

The bulk of the effort on metaphor annotation and computational metaphor identification has focused on English. Most of metaphor annotation projects for Russian which are known to us adhere the conceptual metaphor paradigm, such as the Russian sections of the multilingual resources (Dodge, Hong, & Stickles, 2015; Mohler, Brunson, Rink, & Tomlinson, 2016). The only known to us Russian dataset of linguistic metaphor is the corpus compiled by Tsvetkov, Boytsov, Gershman, Nyberg, & Dyer (2014). However, this dataset in several regards differs from the corpus which was collected and annotated in the present study:

- the corpus by Tsvetkov and colleagues is smaller: its size amounts to a total of 240 sentences, while our corpus comprises more than 7,000 sentences; to the best of our knowledge, this is the largest currently existing Russian corpus annotated for linguistic metaphor;
- the corpus by Tsvetkov and coauthors does not concentrate on any specific set of target lexemes and covers a range of most frequent Russian verbs and adjectives; our corpus is designed around twenty target verbs: this allows us to explore the impact of the linguistic characteristics of verbs on the performance of classification features;
- Tsvetkov et al. report that metaphoric sentences in their corpus were selected so as to contain only one metaphor, that is, the metaphoric occurrence of the target verb or adjective; the corpus presented in this thesis was compiled with the aim of approximating the experimental task to the demands of real-world NLP applications: therefore, it contains sentences which may feature multiple instances of figurative language as well as language errors and inaccuracies.

Next, most of the computational metaphor identification work for Russian that we are aware of also follows the top-down design, i.e. is aimed at identifying conceptual metaphors (Dodge, Hong, & Stickles, 2015; Dunn et al., 2014; Mohler, Rink, Bracewell, & Tomlinson, 2014; Strzalkowski et al., 2013). There are two experiments for identification of linguistic metaphor in Russian that are known

to us: (1) Tsvetkov, Mukomel, and Gershman (2013) and (2) Tsvetkov, Boytsov, Gershman, Nyberg, & Dyer (2014). However, the design of their experiments is substantially different from the experiments conducted in this thesis:

- the experiments by Tsvetkov and coauthors is based on cross-lingual model transfer: classification features in non-English languages are translated into English with a machine-readable dictionary and then they are vectorized using English lexical resources (such as WordNet, the MRC Psycholinguistic Database, or distributional semantic models); our experiments are monolingual: they neither depend on the quality of machine translation which may become problematic in cases of polysemy, nor do they require data from other languages and solely rely on resources that are currently available for Russian NLP.
- the experiments by Tsvetkov and colleagues operate on syntactically related tuples (Adjective-Noun pairs) and triples (Subject-Verb-Object): as a consequence, they are dependent on the quality of syntactic parsing which is not always reliable in real-life tasks; our experiments take full sentential context as input: this enables us to explore the impact of contextual and discourse factors on identification of metaphor.

Scientific novelty of the thesis. As pointed out above, we see the main goal of this thesis in suggesting a linguistic explanation and interpretation of the language and discourse-based factors which promote the success of some computational models of linguistic metaphor identification and cause the other models to falter on the task. The output of a machine learning classifier is analyzed by means of statistical methods and other ML algorithms in order to arrive at empirical, data-driven conclusions about the linguistic mechanisms contributing to metaphor identification in context. To the best of our knowledge, this is the first attempt of such research.

Theoretical significance of the thesis. The findings of the present research may have a value for psycholinguistic and broader cognitive studies. The results presented in the thesis can shed light on the cognitive factors that make processing of metaphor by humans possible, since we explore the lexico-semantic and morphosyntactic cues which are deployed in carrying the signals of metaphoricity across from the speaker to the recipient. The results of this research can help to outline the inventory of metaphor cues and to evaluate their salience. As we look at metaphor in context and apply nonlinear (bag-of-items) representation, it allows us to make conclusions as to whether metaphor can be modelled as a holistic mental process in which the information carried by a verbalized message is a non-compositional unity of its constituent cues. Eventually, the present research may have implications for efforts aimed at providing a computational model of the metaphor decoding and encoding process.

Practical significance of the thesis. The major contributions of this thesis can be summarized as the follows:

– The research re-implemented the approaches to corpus annotation which had been suggested in earlier work on metaphor annotation in English. We introduced minor modification and applied the previously suggested protocols to Russian data;

– We compiled a relatively large dataset of metaphorical and non-metaphorical usages of 20 Russian verbs, which is made available for public use. To the best of our knowledge, this is the first public resource of this kind;

– An annotation validation experiment in a setting with multiple annotators was conducted;

– We release a ranking of concreteness indexes for approximately 17K Russian words;

– The study tested a number of earlier methodologies of feature extraction for metaphor identification in application to Russian (lexical and morphological frequencies, distributional semantic vectors, and concreteness scores);

– We developed a classifier for sentence-level binary-class identification of metaphoric occurrences in raw running Russian text;

– The thesis provides linguistic evaluation of the quality of classification and compares the efficiency of models based on different features;

– We also suggest data-driven linguistic interpretation to the performance of the features and identify the features which hold potential for generalizability;

– The thesis provides analysis aimed at an empirical verification of the theoretical claims that formed the basis of the computational models.

Public demonstrations of the results

The major results of the research were presented at the following events:

- The 2017 Spring Symposium Series of the Association for the Advancement of Artificial Intelligence (Stanford University, Computer Science Department; Palo-Alto, USA, 2017);
- The 2nd Kolmogorov Seminar on Computational Linguistics and Language Studies (National Research University Higher School of Economics, Moscow, Russia, 2017);
- Dialogue-2017, the 23rd International Conference on Computational Linguistics and Intellectual Technologies (Russian State University for the Humanities, Moscow, Russia, 2017);

- RuSSIR-2017, Russian Summer School in Information Retrieval (Ural State University, Yekaterinburg, Russia, 2017);
- The 3nd Kolmogorov Seminar on Computational Linguistics and Language Studies (National Research University Higher School of Economics, Moscow, Russia, 2018);
- AINL-2018, Artificial Intelligence and Natural Language Conference (ITMO University, Saint Petersburg, 2018)
- The 9th International Cognitive Linguistics Congress (National Research University Higher School of Economics, Nizhny Novgorod, 2019).

Note on collaboration

The initial experiments on Russian verbal metaphor identification with distributional semantic features (Panicheva & Badryzlova, 2017b) were led by Polina Panicheva in collaboration with the author of the thesis. All the other theoretical, experimental and composition work involved in the production of the thesis was carried out by the author alone.

Organization of the thesis

The thesis consists of Introduction, four Chapters, Summary, and List of References comprising 206 titles.

Chapter I provides an overview of the state-of-the-art approaches to annotation of metaphor in corpora and to engineering computational systems for automated metaphor identification.

Chapter II is devoted to the experimental corpus – the principles of selecting data and target verbs, and annotating the corpus; the chapter also gives an outline of the metaphoric and non-metaphoric classes and describes the inter-annotator reliability test – the annotator instructions, annotations binarization schemes, and the obtained measure of agreement between the annotators; the last subsection of the chapter looks at the cases of inter-annotator disagreement.

Chapter III details the metaphor identification experiment. It introduces the set of chosen features and explains the theoretical background which motivated the choice. The chapter goes on to describe the statistical approaches and computational resources which were applied in order to convert the input data into vectors, as well as the design of the machine learning experiment. The second half of the chapter discusses the results of the classification experiment: we compare the performance of models and evaluate the utility of increasing the model complexity.

Chapter IV provides in-depth analysis of the linguistic factors determining the performance of the models. We identify the linguistic units which are most probable to carry the signal of metaphoricity and make predictions about their generalizability.

Finally, we present the **Conclusions** of the thesis and make suggestions for future research in the area of computational identification of metaphor.

SUMMARY OF THE THESIS

Both metaphor annotation and computational identification of metaphor can be carried out within one of the two paradigms:

- the conceptual metaphor paradigm (or the "top-down" paradigm). A conceptual metaphor (CM) "consists of two conceptual domains, where one domain [the Target] is understood in terms of another [the Source]" (Kovecses, 2010).
- the linguistic metaphor paradigm (or the "bottom-up" paradigm). A linguistic metaphor (LM) is "a stretch of language that creates the possibility of activating two distinct domains" (Cameron, 2003).

In the conceptual metaphor paradigm, since the annotator of a corpus or the designer of a computational system presume the presence of a conceptual metaphor, they start out either from a set of predefined conceptual metaphors or from particular source or target domains and then search for linguistic expressions that are compatible with them.

In the linguistic metaphor paradigm, the annotator and the system designer presume no specific conceptual metaphor; they search for any lexical units that are used indirectly, and only at a later stage mappings can be derived from the linguistic expressions that have been identified as metaphorically used.

Chapter I explains the difference between the two paradigms and gives a brief outline of the corpora and systems designed within the conceptual metaphor paradigm. However, major emphasis is given to the projects carried out within the linguistic metaphor paradigm. This is due to the focus of the present thesis: the experimental corpus presented in Chapter II is annotated in the bottom-up manner, and the experiment for computational metaphor identification in Russian in Chapter III is designed within the linguistic metaphor paradigm.

Some of the **major conceptual metaphor annotation projects**, are listed below:

- The LCC Metaphor Dataset (Mohler, Brunson, Rink, & Tomlinson, 2016) is available for English, Russian, Spanish, and Farsi;
- MetaNet (Dodge, Hong, & Stickles, 2015) is also a multi-lingual repository of CMs in English, Spanish, and Russian;
- The corpus by (Shutova & Teufel, 2010) is available for English, it contains two layers of annotation, one for LMs and another for CMs.

As for the **resources annotated for linguistic metaphor**, Section 1 of Chapter I focuses on the largest annotated corpus of linguistic metaphor, the VU Amsterdam Metaphor Corpus, or VUAMC, and on the metaphor identification procedure (MIPVU) which was designed to annotate VUAMC (G. J. Steen et al., 2010).

VUAMC contains sentences from the four subdomains of the British National Corpus (BNC) – academic, news, fiction, and conversation – with a total of approx. 200,000 words. The corpus contains annotations of only LMs of the following types: indirect metaphor-related words (MRWs), direct MRWs, implicit MRWs (i.e. pronominal antecedents of metaphorically expressed referents), possible personification, metaphor flags (lexical signals of metaphor), and borderline cases (WIDLII). All annotations were performed manually on a word-to-word basis by five expert linguists. VUAMC was used as the training and test dataset in the first shared metaphor detection task (Leong, Klebanov, & Shutova, 2018).

The steps of the Metaphor Identification Procedure (MIPVU) can briefly be summarized as follows:

- Identify the contextual meaning of a lexical unit;
- Check if there is a more basic meaning of the lexical unit. The basic meaning is a more concrete, specific, and human-oriented sense in contemporary language use;
- Determine whether the more basic meaning is sufficiently distinct from the contextual meaning;
- Examine whether the contextual meaning can be related to the more basic meaning by some form of analogy (Steen et al., 2010, pp. 33–35)

For instances of computational metaphor identification system designed within the conceptual metaphor paradigm, consider the works of (Bracewell, Tomlinson, Mohler, & Rink, 2014; Gedigian, Bryant, Narayanan, & Ciric, 2006; Heintz et al., 2013; Mohler, Bracewell, Hinote, & Tomlinson, 2013; Mohler, Rink, Bracewell, & Tomlinson, 2014; Strzalkowski et al., 2013).

Computational systems for identification of linguistic metaphor are exemplified by the following projects: (Bulat, Clark, & Shutova, 2017; Hovy et al., 2013; Klebanov, Leong, Heilman, & Flor, 2014;

Klebanov, Leong, Gutierrez, Shutova, & Flor, 2016; Mu, Yannakoudakis, & Shutova, 2019; Shutova, Kiela, & Maillard, 2016; Stemle & Onysko, 2018; Turney, Neuman, Assaf, & Cohen, 2011; Wu et al., 2018).

Besides the differences in the paradigm (CM vs. LM), experiments in computational identification of metaphor are differentiated by the settings in which they are designed -- **supervised, unsupervised, or deep learning**.

Since the experiments for Russian metaphor identification described in Chapters II and III of the present thesis follow the paradigm of linguistic metaphor and are designed in the supervised binary classification setting, Chapter I devotes major attention to description of projects which are similar to ours in their design.

Computational systems for metaphor identification also differ in **the types of features** exploited in them:

- Lexical features (e.g. Klebanov, Leong, Heilman, & Flor, 2014);
- Morphological and syntactic features (e.g. Hovy et al., 2013; Ovchinnikova et al., 2014);
- Distributional semantic features (e.g. Shutova, Kiela, & Maillard, 2016);
- Topic modelling (e.g. Heintz et al., 2013);
- Lexical thesauri and ontologies: WordNet (e.g. Gandy et al., 2013), FrameNet (e.g. Gedigian, Bryant, Narayanan, & Ciric, 2006), VerbNet (e.g. Klebanov, Leong, et al., 2016), ConceptNet (Ovchinnikova et al., 2014), and the SUMO ontology (Dunn, 2013a, 2013b);
- Psycholinguistic features: concretness / abstractness, imageability, affect, and force (e.g. Neuman et al., 2013; Strzalkowski et al., 2013; Turney et al., 2011).

Results of metaphor identification experiments are difficult to compare for a number of reasons: (a) the theoretical incompatibility and the subsequent differences in the experimental design; (b) some systems identify metaphors on the sentence level while others identify word-level metaphors; (c) many of the existing systems are domain-specific; and (d) most systems are trained and evaluated on different datasets.

Metaphor identification in Russian texts has been addressed in several projects. For example, (Mohler, Rink, Bracewell, & Tomlinson, 2014; Ovchinnikova et al., 2014; Strzalkowski et al., 2013) use a variety of features to model the conceptual source and target domains and to align them with their linguistic realizations in text. (Tsvetkov, Boytsov, Gershman, Nyberg, & Dyer, 2014) and (Panicheva & Badryzlova, 2017a) operate outside of the conceptual metaphor paradigm. The former system exploits cross-linguistic metaphors: the classifier is first trained on the English data, and then the

trained model is projected to Russian using a dictionary. The latter project uses distributional semantic vectors to distinguish metaphoric and non-metaphoric sentences.

Chapter II gives an account of the Russian corpus annotated for metaphoricity, which was used in the subsequent machine learning experiments for identification of metaphor.

The corpus consists of approximately 7,000 sentences which were selected from the large web-crawled corpus ruTenTen11[1]. Each sentence contains an occurrence of one of the 20 target verbs presented in Table 1.

Table 1. Target verbs and number of sentences in the corpus

Russian	**Transliteration**	**Translation (primary meaning)**	**# of sentences**
бомбардировать	bombardirovat	to bombard (smth/smb)	287
доить	doit	to milk (e.g. a cow)	467
греть	gret	to heat / warm (smb / smth)	503
нападать	napadat	to attack (smth/smb)	313
очерчивать	ocherchivat	to outline (smth)	225
отрубать	otrubat	to hack (smth) off	377
пилить	pilit	to saw (smth)	310
подхватывать	podkhvatyvat	to catch (smth falling)	373
причесывать	prichesyvat	to comb (smth/smb)	400
распылять	raspylyat	to spray (smth)	285
разбавлять	razbavlyat	to dilute, to liquefy (smth)	289
съедать	syedat	to eat (smth) up	693
трубить	trubit	to blow a trumpet	397
укалывать	ukolot	to prick (smth/smb)	353
утюжить	utyuzhit	to iron (clothes)	364
выкраивать	vykraivat	to cut (in sewing:	253

[1] accessed via the SketchEngine interface (Kilgarriff et al., 2014)

		parts of a garment, from fabric)	
взрывать	vzryvat	to blow (smth) up, to explode (smth)	289
взвешивать	vzveshivat	to weigh (smth)	298
зажигать	zazhigat	to ignite (smth)	294
жонглировать	zhonglirovat	to juggle (smth)	396
Total: 7,166			

Each sentence in the corpus is annotated as either metaphoric (MET) -- when the target verb is used metaphorically, or as non-metaphoric (NONMET) -- when the target verb is used as non-metaphor. The subcorpus of each target verb is balanced by class: 50% of the sentences are metaphoric, and 50% are non-metaphoric.

When making judgements about metaphoricity of verbs we largely followed the lines of the MIPVU procedure set out in Chapter I, with a number of slight modifications. Thus, we re-interpret the notion of MIPVU's **basic meaning** in terms of its argument structure and use the term **central literal meaning** instead. A central literal meaning is the meaning which:

- authorises a two-actant construction with the following mandatory arguments: (1) the Agent, (2) the Patient / the Theme;
- the Agent of the meaning in question denotes a human being(s); the other arguments refer to physical (concrete, non-abstract) entities.

Since our experimental corpus is collected for the purpose of binary classification, our annotation scheme contains no tag for borderline cases ('WIDLII', in the terminology of MIPVU). Besides, we make no fine-grained distinction between the subtypes of MRWs (metaphor-related word) suggested in MIPVU (such as direct MRWs, indirect MRWs, and possible personification): all these subtypes are conflated into one metaphoric class (MET). Moreover, we avoid using the terms 'indirect MRW' and 'direct MRW', and choose to use the terms 'conventionalized metaphor' and 'unconventional metaphor' instead.

The non-metaphoric class of the corpus comprises contexts where the target verbs are used either in the central literal meaning (Example [1]), or in the meanings that are related to the central literal meaning via either a diathetic shift (i.e. the change of the syntactic rank of the actants), or a close metonymic shift (Example [2]).

[1] (NONMET) *После того , как вы уже < очертили > карандашом контур , слегка припудрите губки…* ‘After you have < outlined > the lips with a pencil, dab some powder over them.’

[2] (NONMET) *Для этого лучше использовать угольный карандаш . Он четко < очертит > мелкие детали и придаст картине законченность и филигранность .* ‘Charcoal pencil is the most suitable for this kind of work. It will < outline > minute details with much precision and will impart to the artwork an appearance of perfect finish and detailedness.’

The metaphoric class of the corpus is represented by the following subtypes of contexts:

A. Conventionalized metaphors (Example [3]);
B. Unconventional metaphors (Example [4]);
C. Idiomatic expressions (Example [5]).

[3] (MET) *СМИ < трубят > о достижениях в решении различных социально-экономических проблем.* ‘Media outlets < trumpet > the progress in solving various social and economic problems.’

[4] (MET) *Самолюбие – это наполненный ветром воздушный шар , из которого вырывается буря , лишь < уколешь > его .* ‘Vanity is a balloon filled with the wind; once you < prick > it, you release a storm.’

[5] (MET) *Когда то мои пра - пра - пра - пра - прадеды … < грели > руки на ростовщичестве.* ‘There was a time when my fore- fore-fore-fore-fore-forefathers used to < warm > their hands (*lit.* to make dishonest or illegal profit) by usury.’

In order to assess the reliability of annotation we carried out a reliability test in which three analysts independently annotated 20% of sentences from the subcorpus of each verb. The sentences for the reliability test were selected so as to equally represent cases that had been deemed problematic for human judgement and cases that were found unproblematic. The test yielded high coefficients of inter-annotator agreement: 0.83 and 0.9 (depending on the scheme that was used to binarize the annotations performed in terms of categorial classes, see Section 3 of Chapter II).

Chapter III offers an in-depth description of the experiment for identification of metaphor in Russian texts on the basis of the experimental corpus introduced in Chapter II. As we pointed out earlier, the experiment is set in the paradigm of linguistic metaphor identification; we train a binary classifier to differentiate sentences containing metaphoric occurrences of target verbs from sentences where the target verbs are used non-metaphorically. We use the Support Vector Machine (SVM) classifier with linear kernel; the training and testing are done with 5-fold cross-validation.

In order to inform the classifier, we extract the following types of features.

1. Distributional semantic vectors; they are computed using a pre-trained word-embeddings model (Kutuzov & Kuzmenko, 2016; 'RusVectōrēs', n.d.) which was trained with word2vec's Continuous Skip Gram algorithm on the 10bn Araneum web corpus (Benko & Zakharov, 2016). We apply the semantic similarity measure (Herbelot & Kochmar, 2016; Newman et al., 2010) which is intended to capture the linear semantic deviances in the text.
2. Lexical co-occurrence vectors; this feature is computed by means of the ΔP metric (Levshina, 2015); the obtained scores show the metaphor association indexes of lexemes in the corpus.
3. Morphosyntactic co-occurrence vectors are computed using the same ΔP measure on full morphological tags of nouns and verbs (while all the other parts of speech are represented only with their part-of-speech tags, and punctuation marks are represented by their lemmas). The resulting indexes demonstrate the association between grammatical categories and metaphor (therefore, we refer to them as indexes of morphosyntactic metaphor association).
4. Indexes of concreteness are computed using the seed list of approx. 360 concrete ('thingness') paradigm nouns and an equal number of abstract paradigm nouns. For each word of the corpus, we measure its semantic similarity to the ten semantically nearest nouns from the thingness paradigm and take the mean of these similarities.
5. Flag words and quotation vectors are computed as the number of their occurrences and the linear distance (in tokens) to the target verb. These latter features proved inefficient for the classification task due to the sparsity of the data; however, we presume that they can be used in future experiments in order to apply weighting schemes to lexical and morphosyntactic co-occurrence vectors.

Results of the metaphor classification experiment (accuracy) for uni-feature models are summarized in Table 2.

Table 2. Results of the metaphor classification experiment (accuracy), uni-feature models (*sem* – distributional semantic *lex* – lexical co-occurrence, *morph* – morphosyntactic co-occurrence, *concr* – concreteness)

datasets/models	sem	lex	morph	concr_abstr
bombardirovat	0.75	0.82	0.74	0.71
doit	0.69	0.81	0.74	0.77
gret	0.69	0.87	0.7	0.85
napadat	0.58	0.75	0.73	0.62
ocherchivat	0.6	0.84	0.72	0.91
otrubat	0.64	0.84	0.75	0.65
pilit	0.55	0.8	0.74	0.8

podkhvatyvat	0.68	0.83	0.75	0.79
prichesyvat	0.72	0.91	0.77	0.83
raspylyat	0.8	0.91	0.78	0.87
razbavlyat	0.78	0.88	0.76	0.81
syedat	0.8	0.85	0.73	0.74
trubit	0.79	0.81	0.75	0.79
ukalyvat	0.51	0.78	0.72	0.77
utyuzhit	0.66	0.86	0.73	0.7
vykraivat	0.79	0.94	0.84	0.86
vzryvat	0.52	0.84	0.69	0.69
vzveshivat	0.57	0.82	0.73	0.85
zazhigat	0.63	0.85	0.69	0.76
zhonglirovat	0.66	0.78	0.7	0.83
combined dataset	0.65	0.82	0.67	0.76
mean of 20 datasets	0.67	0.84	0.74	0.78
std of 20 datasets	0.095	0.047	0.034	0.076

We report that the model based on lexical co-occurrences appears to hold the greatest potential for generalizability and for being used on new unseen data: this feature both performs consistently well across the individual datasets of the 20 target verbs and yields a high accuracy (0.82) on the combined dataset of the 20 verbs. The morphosyntactic model shows consistently lower results across the 20 individual datasets and falls behind when applied to the combined dataset, yielding the accuracy of 0.67. Although the distributional semantic vectors behave consistently, their performance is always low – both on individual datasets and on the combined dataset. The behavior of the concreteness feature is highly inconsistent: it produces very high accuracy (0.9-0.91) on some verbs while sinking to the chance level on the others; the accuracy on the combined dataset is the lowest of all the four models – 0.63.

Different models deliver different quality of classification across the datasets, as well as the datasets vary in their performance across the models. Some of the datasets yield high accuracy with one or two single-feature models, and fail with the others; other datasets, however, perform equally well with all or most of the models. Presumably, these differences should be attributed to the divergent patterns of semantic, lexical, and morphological combinability evoked by different verbs and their different meanings – the issue which is addressed in Chapter IV.

On 19 of the 20 individual datasets, as well as on the combined dataset, the optimal performance is achieved with one-feature models; further increasing the complexity of the model by applying

additional features occasionally leads to a slight increase (by 1-2 percentage points), but the tradeoff between the gain and the complexity (i.e. the utility of the gain) is not convincing.

The robustness of the lexical co-occurrence feature across the 20 datasets, beside holding promise for generalizability, also means that certain lexemes can function as predictors of metaphoricity which can be expected to persist when applied to new unseen data. An attempt to identify such lexical predictors is to be made in Chapter IV.

Although the distributional semantic, morphological co-occurrence, and concreteness features did not live up to expectations, a more in-depth analysis of their performance is likely to reveal valuable insights into the semantic and morphological factors of metaphoricity; this analysis is conducted in Chapter IV.

Section 1 of Chapter IV discusses the results of the lexical and the concreteness classifiers which are analyzed in conjunction with each other.

The features implemented in the lexical classifier were analyzed, firstly, in terms of the metaphor association indexes of lexemes, and secondly, in terms of the importance of lexemes. The metaphor association indexes are computed as the ΔP coefficients introduced in Chapter III. The importance of lexemes is evaluated automatically by means of the feature selection functionality available in the Random Forest classification algorithm.

The features utilized in the concreteness classifier were analyzed in terms of the concreteness scores of lexemes, which were computed as described in Chapter III.

We hypothesized that the accuracy of classification may be correlated with the degree to which the metaphor-specific and the non-metaphor-specific vocabularies are juxtaposed to each other, i.e. with the sizes of these two groups. In order to assess the sizes of metaphor-specific and non-metaphor-specific lexicons, we clustered the metaphor association indexes of the lexemes in the subcorpora of each target verb; we observed that target verbs differ in the extent of words that exclusively associate either with metaphor or non-metaphor. In order to establish whether the efficiency of classification correlates with the size of metaphoric and non-metaphoric vocabulary, we computed Pearson correlation between these two variables. As a result, we obtained moderate correlation between the accuracy of classification and the size of both non-metaphoric vocabulary (correlation =0.55), and metaphoric vocabulary (correlation=0.44). This allows us to conclude that classifiers operating on lexical features should perform better on datasets where two distinct groups of lexis are present: (a) words which are specific to metaphoric contexts and (b) words that predominantly occur in non-metaphoric contexts.

Next, we analyzed several cases of lexical features which were ranked as important by the feature selection algorithm in datasets of individual target verbs. The analysis lead us to conclusion that the automatically computed importance of lexical features reflects the patterns of the verbs' lexico-semantic combinability: for example, for the verb *vykraivat*, the most important lexemes with high metaphor association are *время, час, день, деньги, график, бюджет, неделя, пара, мир, отпуск*, while the most important lexemes with the low metaphor association are *припуск, сторона, срез, кожа, сумка, вариант, клапан, заготовка, рулон, отрезок*.

As we showed in Chapter III, of all the types of features tested in our classification experiment, lexical features appear to be the most promising in terms of generalizability, i.e. the potential to yield reasonable accuracy when the classifier is trained on our experimental dataset and then tested on new unseen data. This optimism is motivated by the high accuracy (0.82) demonstrated by the lexical classifier on the combined dataset of the 20 target verbs; the high accuracy achieved when the classifier traverses across the target verbs suggests that there should exist a set of vocabulary which is reliably ported across the idiosyncratic lexico-semantic combinability patterns of individual verbs, or, to put it differently, there is a group of words that are shared by all metaphoric contexts, on the one hand, and by all non-metaphoric contexts, on the other -- irrespective of the target verb.

We presume that the lexemes that may be expected to generalize well and to act as predictors for classification on new unseen data are likely to be found among the two sets of lexical items: (a) lexemes that were estimated to be the most important features on the combined dataset of the 20 verbs and (b) lexemes with the highest average importance across the 20 individual target verb datasets. In order to pinpoint the predictors we concatenate the two lists and filter them by the variance of their distribution across the individual target verb datasets (thus removing infrequent words and lexemes that are attracted by a small number of verbs). The resulting list of potential lexical predictors is presented in Table 6 in Chapter IV. For example, nominal predictors with high indexes of metaphor association are *задача, итог, риск, позиция, необходимость, уровень, житель, регион, факт, понятие, процесс, проблема, качество, состояние, центр, оценка, цель, результат, условие, период, отношение, развитие*, etc.; non-metaphoric predictors expressed by nouns are *платье, сумка, рукав , топор, нож, лампа, литр, фонарь, нить, батарея, телефон, мл, шерсть, торт, костер, песня, каша, жидкость, камень, чеснок, факел, девушка, лошадь, книга*, etc. We observe that the list of possible lexical predictors contains a substantially larger number of nouns than verbs, adjectives, or adverbs; consequently, we conclude that nouns hold greater potential for metaphor classification with lexical features than other parts of speech.

The validity of the suggested list of potential lexical predictors requires further empirical testing; in order to see how many of them persist, a corpus of new target verbs should be collected replicating the design of our experimental dataset. Then, potential lexical predictors should be extracted from the new corpus with the same method as was implemented in Chapter IV (Random Forest feature selection, filtering by the variance of distribution thresholds). Overlap between the present list of predictors and the new list of predictors will indicate the most stable and generalizable predictors for classification of metaphor with lexical features.

Next, we scrutinized how the degree of metaphor association of lexemes correlates with their concreteness. Firstly, we computed the average concreteness of the three groups of words in the combined dataset of the 20 target verbs: (a) words with high metaphor association indexes, (b) words with medium association indexes, and (c) words with low association indexes. Secondly, we conducted a pairwise independent-sample t-test to compare the concreteness of the three clusters. The concreteness and metaphor association are inversely related: the cluster with the highest metaphor association has the lowest average concreteness, and the cluster with the lowest metaphor association has the highest average concreteness. The t-test confirmed that the differences in concreteness are statistically significant across the clusters. The observation that words which are strongly associated with metaphor tend to be less concrete than words that are associated with non-metaphor seems to be of high value to metaphor studies: it lends empirical support to a central premise of the conceptual metaphor theory which postulates metaphor as a mapping between concrete source domains and abstract target domains.

Section 2 of Chapter IV is devoted to qualitative analysis of the performance of the classifier based on distributional semantic features. As we showed in Chapter III, the performance of this model is inconsistent across the datasets of individual target verbs (with the accuracies of classification ranging from 0.5 to 0.82, mean=0.67), and the accuracy on the combined dataset (0.65) is much lower than the corresponding accuracy of the classifier based on lexical unigrams (0.82). The verbs which yield the highest accuracy with the distributional semantic classifier are *syedat* ‘to eat smth up’, *vykraivat* ‘to cut (in sewing: parts of a garment, from fabric)’, *raspylyat*, ‘to spray smth’, and *razbavlyat* ‘to dilute, to liquefy (smth)’; the verbs with the lowest accuracy are *vzveshivat* ‘to weigh smb/smth’ and *vzorvat* ‘to blow smb/smth up’.

Having analyzed the metaphoric and the non-metaphoric contexts of these six verbs we arrive at the conclusion that the success of classification very likely depends on the semantic homogeneity of the metaphoric and / or the non-metaphoric subcorpora of each verb: in most successful verbs, at least one of the subcorpora is semantically homogeneous, and the two subcorpora are distinctly juxtaposed

to each other, i.e. there are no semantic classes which are shared by metaphoric and non-metaphoric sentences. In addition, the ratio of words which are low in polysemy (i.e. which are either monosemous or have a small number of meanings) may play a role, since the pre-trained distributional semantic models from which we obtain the vectors for our classification do not disambiguate the meanings of polysemous verbs; this is likely to negatively impact the accuracy of classification.

For example, practically all non-metaphoric sentences with the verb *syedat* 'to eat smth up' contain terms from the semantic field FOOD, and these lexemes are predominantly low-polysemous (e.g. *йогурт, инжир, завтрак, овощи, мороженое*, *салат, конфеты, каша, суп, ананас, омлет*, etc.). Most of the non-metaphoric sentences with *raspylyat* 'to spay smth' and *razbavlyat* 'to dilute, to liquefy (smth)' contain (predominantly) low-polysemous lexemes from the semantic fields of liquids and granulated, gaseous, or powder-like substances (e.g. *вода* 'water', *аэрозоль, спрей* 'spray', *химикаты* 'chemicals', *газ* 'gas', *краска* 'paint', *раствор*, *препарат* 'solution', etc.). The verb *vykraivat* '' to cut (in sewing: parts of a garment, from fabric)' is remarkable in that its both the metaphoric and the non-metaphoric subcorpora seem to possess a high degree of semantic homogeneity: while the non-metaphoric sentences abound in terms from the domain of clothes-making (e.g. *воротник* 'collar', *карман* 'pocket', *подкладка* 'lining', *ткань* 'fabric', *рукав* 'sleeve', *шов* 'seam', *пояс* 'belt', *трикотаж* 'knitted fabric', *обтачка*, *бейка* 'piping', etc.), the non-metaphoric contexts tend to contain vocabulary from the three semantic fields, TIME, MONEY, and SPACE.

As for the factors that contribute to the poor performance of the distributional semantic classifier on the verb *vzveshivat* 'to weigh smb/smth', it seems to be conditioned by the semantic vagueness of the direct object of its non-metaphoric meaning – since the action of weighing can be applied to virtually any physical object. Besides, the semantic borderline between the metaphoric and non-metaphoric subcorpora of *vzveshivat* appears to be fuzzy, since either of them contains a significant proportion of highly abstract lexemes. In the case of non-metaphoric sentences, their occurrences are triggered by the parametric semantic of the verb (such abstract words as *количество* 'amount / quantity', *часть* 'part', *вещество* 'substance', *тяжелый* 'heavy', *легкий* 'light', *обнаружить* 'to identify', *сосчитать* 'to count', *рассчитать / посчитать* 'to calculate', *получить* 'to obtain', *узнать* 'to find out', *измерить* 'to measure'; *изменение* 'change / modification', *состояние* 'state / condition', *концентрация* 'concentration'; *точность* 'precision', *процент* 'per cent', *цифра* 'figure / number', *функция* 'function'; *полезный* 'useful / effective', *аккуратно* 'carefully', *конкретный* 'concrete', etc.).

In the case of the verb *vzorvat* 'to blow smb/smth up', the performance of the classifier near the chance level could be attributed to the following two factors both of which act towards blurring the borderline between the metaphoric and the non-metaphoric subcorpora. Firstly, the direct objects of one of its metaphoric meaning (insert the meaning) tend to be expressed metonymically, e.g.: *Теперь оно (правительство) хочет < взорвать > школу изнутри – социальным неравенством* . 'And now it (the government) wants to < blow up > school from inside – with social inequality.' Secondly, the other metaphoric meaning of *vzorvat* tends to be used in discourse describing war- and terrorism-related contexts, e.g.: *Из Кабардино-Балкарии можно < взорвать > ситуацию на всем Северном Кавказе* . The situation in the entire North Caucasus can be < blown up > from [*Russia's North Caucasus Republic of*] Kabardino-Balkaria.

Section 3 of Chapter IV starts by ascertaining that the accuracy of classification depends on the presence of salient morphosyntactic patterns licensed by a target verb. Their presence is manifested in the degree to which the grammatical categories with the highest indexes of metaphor association are juxtaposed to the grammatical categories with the lowest indexes. This degree can be visualized and quantified by measuring the slope and the variance of the curve composed of the metaphor association indexes of the morphosyntactic categories found in the dataset of each target verb. We found strong Pearson correlation between the accuracy of classification and the slopes and variances of curves of the 20 individual target verbs (-0.75 and 0.78 correspondingly).

Section 3 goes on to analyze the importance of individual morphosyntactic categories for the classification task. The importance of morphosyntactic features is evaluated by means of the Boruta algorithm (Kursa, Jankowski, & Rudnicki, 2010), which is an extension of the Random Forest method returning the decision for each feature ('Confirmed', 'Tentative', or 'Rejected').

We start by looking at the features that are found important on the dataset of one specific verb, *vykraivat* 'to cut (in sewing: parts of a garment, from fabric)'; we demonstrate how the important morphosyntactic categories of this verb are realized in metaphoric and non-metaphoric contexts. For example, the category *n* (the neuter gender of nouns, adjectives, and past verbs) gains importance due to the frequent occurrence of abstract neuter nouns and nominal pronouns in the position of the direct object when the verb is used metaphorically[2] (such neuter forms as *время*, *то*, *это*, *посещение*, *дело*, *занятие*, *мероприятие*, etc.). We show that the frequent occurrence of these forms is licensed by the

[2] 'Possessing a limited amount of resource A3, person A1 makes an effort to preserve a smaller part A2 of this resource in order to use it for action A4 or for an action related to A4'.(Apresyan et al., 2014)

following construction: (modal expression) ***выкроить*** *время* (time / period of time) на (для)_Prp N_acc / на то (для того), чтобы_Conj V_inf.

We go on to analyze the grammatical categories that were confirmed to be relevant ten or more times across the 20 datasets of the individual verbs, i.e. the categories that can be regarded as reliable predictors that are common to all the verbs. We zoom in at each of these categories and demonstrate how they are realized in discourse. For instance, the category *part* (particle) is important and strongly associated with metaphor. We divide all particles into three groups: (a) particles expressing negation (*не* and *ни*), (b) the two variants of the particle used to express the conditional aspectual forms of verbs (*бы* and *б*), and (c) discourse particles used to mark the sentential focus and to serve the rhetoric function (e.g. *же* , *только* , *даже* , *просто* , *вот* , *лишь* , *ли* , *именно*, etc.). We show that particles of all the three groups occur substantially more frequently in metaphoric than in non-metaphoric contexts, i.e. language users tend to use more negations, conditionals, and discourse markers in metaphoric sentences. We conjecture that this bias may be interrelated with the genre and domain specifics of the metaphoric and non-metaphoric contexts in our corpus: while the latter contain more instructional and technical discourse, the former contain more sentences that were retrieved from web fora, social media and the like, where users write in a language which approximates spoken discourse; spoken discourse, in its turn, is construed with the aim of conveying an immediate and often personal message to the other person -- and thus it is characterized by more intensive expression of emotions and shades of modality.

High importance is assigned by the Boruta algorithm to the metaphor-associated grammatical category *n* (the neuter gender of nouns, adjectives, and past verbs) -- the behavior that we have already observed in the individual case of the verb *vykraivat* above and which was accounted for by high frequency of abstract neuter forms. We demonstrate that this observation holds true across the target verbs: metaphoric sentences tend to feature more neuter forms than non-metaphoric contexts, and this is related with their greater abstractness. We compute the mean concreteness of neuter vs. feminine vs. masculine adjective and nouns in the entire experimental corpus using the concreteness scores introduced in Chapter III. Neuter forms possess lower concreteness, and we apply pairwise T-statistic to demonstrate that the difference in concreteness for neuter vs. feminine and neuter vs. masculine is statistically much more significant than for feminine vs. masculine.

Finally, we analyze the grammatical categories that are identified as being important for classification on the combined dataset of the 20 target verbs. We observe that the important grammatical categories

of the individual verbs' datasets and the important grammatical categories of the combined dataset are mirror opposites of each other: the categories that are found to be important in one case are unimportant in the other, and vice versa. We presume that this controversial behavior corroborates low generalizability of the morphosyntactic feature, which was demonstrated in Section 5 of Chapter III. We predict that in an experiment where the classifier will be trained on the morphosyntactic features extracted from our experimental corpus and then tested on morphosyntactic features from a new unseen dataset, the accuracy of classification will be low.

The Conclusions sum up the major contributions of the research:

- it demonstrated implementation of the state-of-the-art approaches to automatic metaphor identification on Russian data;
- suggested new approaches to feature engineering for automatic metaphor identification;
- the insights into the nature of the factors of metaphoricity gained due to the in-depth data-driven analysis of the experimental results may help to improve and advance computational methods for automatic metaphor identification.

The following **papers have been published** on the topic of the present thesis; three papers are devoted to computational metaphor identification and two to annotation of metaphor in corpus:

1. Badryzlova, Y. (2017). Opy`t korpusnogo modelirovaniya faktorov metaforichnosti na primere russkix glagolov [A corpus-based study of factors and models of metaphoricity: evidence from Russian verbs]. *Computational Linguistics and Intellectual Technologies*, *2*, 30–44. Moscow.
2. Badryzlova, Y., & Lyashevskaya, O. (2017). Metaphor Shifts in Constructions: The Russian Metaphor Corpus. *The 2017 AAAI Spring Symposium Series: Technical Reports*, 127–130.
3. Badryzlova, Y., Lyashevskaya, O., & Panicheva, P. (2019). Computer and metaphor: when lexicon, morphology, punctuation, and other beasts fail to predict sentence metaphoricity. *Cognitive Studies of Language. Integrative Processes in Cognitive Linguistics*, *37*, 609–615. Nizhny Novgorod.
4. Badryzlova, Y., & Panicheva, P. (2018). A Multi-feature Classifier for Verbal Metaphor Identification in Russian Texts. *Conference on Artificial Intelligence and Natural Language*, 23–34. Springer.
5. Panicheva, P., & Badryzlova, Y. (2017). Distributional semantic features in Russian verbal metaphor identification. *Computational Linguistics and Intellectual Technologies*, *1*, 179–190. Moscow.