NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

as a manuscript

Yulia Badryzlova

AUTOMATED METAPHOR IDENTIFICATION IN RUSSIAN TEXTS

Dissertation for the degree
of Candidate of Science in Philology and Linguistics HSE

Moscow 2019

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»»

На правах рукописи

Юлия Геннадьевна Бадрызлова

АВТОМАТИЧЕСКИЕ МЕТОДЫ РАСПОЗНАВАНИЯ МЕТАФОРЫ В ТЕКСТАХ НА РУССКОМ ЯЗЫКЕ

Специальность 10.02.20 «Прикладная и математическая лингвистика»

Диссертация на соискание ученой степени кандидата филологических наук
НИУ ВШЭ

Москва 2019

Table of Contents

Chapter I. Metaphor as a computational problem

1. Annotated corpora and datasets

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 (G. J. 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 (G. J. Steen et al., 2010). A comprehensive overview of state-of-the-art approaches to automated metaphor identification is available in (Veale, Shutova, & Klebanov, 2016).

Given such magnitude on computational effort on metaphor, the role of annotated datasets for training and testing and NLP systems becomes paramount.

The VU Amsterdam Metaphor Corpus, VUAMC (G. J. Steen et al., 2010), is the largest and the most widely used dataset. It contains sentences from the four subdomains of the British National Corpus (BNC) – academic, news, fiction, and conversation – with a total of approx. 200K words. The corpus contains annotations of only LMs of the following types: indirect (~conventionalized) metaphor, direct (~novel, unconventional) metaphor, implicit metaphor (~pronominal antecedents of metaphorically expressed referents), personification, metaphor flags (words signaling metaphor) and borderline cases. 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.

The Metaphor Identification Procedure (MIPVU) which was used to define the metaphoricity of lexical units when annotating VUAMC has gained much popularity. In brief, it can be expressed as the following:

- 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 (G. J. Steen et al., 2010, pp. 33–35)

The corpus by (Shutova & Teufel, 2010) contains two layers of annotation, one for LMs and another for CMs; only verbs are annotated. The sentences are taken from the BNC, representing a wide range of genres (literature, newspaper/journal articles, essays on politics, international relations and history, and radio broadcasts). The size of the resulting corpus is 761 sentences which were annotated by three experts independently. When annotating LMs, the annotators followed by guidelines of MIP (an earlier version of MIPVU, see (Pragglejaz Group, 2007). When identifying Source-Target mappings, the annotators drew on the Master Metaphor List (Lakoff, Espenson, & Schwartz, 1991) and extended it with novel categories.

MetaNet (Dodge, Hong, & Stickles, 2015) is a large structured hand-crafted multi-lingual repository of conceptual metaphors; it is reported to contain 787 CMs in English, 547 CMs in Spanish, and 127 Russian CMs. There are CMs of two types: a) CM representing well-studied metaphoric schemata (of various degrees of compositionality and complexity), such as MORE IS UP or MIND IS A MACHINE, and b) domain-specific CMs from the spheres of political and social issues (Government, Wealth, Crime, and others), e.g. GOVERNMENT IS A PARENT or CRIME IS A DISEASE. For each CM, the Source and the Target domains are identified; their semantic structure is presented in terms of frames (i.e. schematic representations of different kinds of experiences, objects, and events), and frame-to-frame mappings between the Source and the Target are provided. The CMs and frames in the repository are organized in a hierarchical structure. Each frame includes a list of lexical units that evoke it. The CMs also contain entailments (inferences that can be made from the frame structure of a CM, as well as entailed CMs). Each CM is illustrated by several sentences.

The LCC Metaphor Dataset (Mohler, Brunson, Rink, & Tomlinson, 2016) is a multi-lingual resource of CMs which is comprised of sentences where the tenor (the lexical units expressing the Source) and the vehicle (the lexical units expressing the Target) are syntactically related to each other. Each of the valid sentences was manually ranked by experts on the three aspects of the tenor / vehicle pair: 1) metaphoricity score (on a four-point scale) 2) mappings between tenors and Sources/vehicles and Targets and 3) affect (emotional impact). The dataset also contains negative material (i.e. non-metaphoric and syntactically unrelated occurrences of the tenor/vehicle pairs). The English subset of is reported to contain a total of 107.5K metaphoric/54K negative (non-metaphoric + non-syntactic) occurrences, the Spanish subset – approx. 64K metaphoric/42K negative pairs, the Russian subset – about 28K metaphors/32K negative occurrences, and the Farsi part – approx. 61K metaphoric/42K negative pairs.

2. Computational modeling of metaphor

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.

A linguistic metaphor (LM) is "a stretch of language that creates the possibility of activating two distinct domains" (Cameron, 2003). NLP systems designed within this paradigm aim to identify any stretches of text that contain indirectly used lexical units: (Klebanov, Leong, Gutierrez, Shutova, & Flor, 2016; Neuman et al., 2013; Wu et al., 2018). Such systems can be integrated in more complex architectures for identification of CMs as the first stage which generates candidates for the subsequent CM analyzer.

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 NLP systems that adhere to this framework aim to detect realizations of conceptual metaphors in the surface structure of text. There are two types of such systems a) those following the top-down design, i.e. they proceed from a set of predefined CMs and search for lexemes that fill the Source and Target slots (Dodge et al., 2015; Rosen, 2018) and b) those that attempt to induce Source-Target mappings from a corpus in an unsupervised or weakly supervised manner (Shutova, Sun, Gutiérrez, Lichtenstein, & Narayanan, 2017).

Systems for metaphor identification are designed in the supervised and the unsupervised, as well as neural networks settings.

The following types of text-internal and text-external features are exploited for metaphor identification:

- 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 and imageability (Neuman et al., 2013; e.g. Turney, Neuman, Assaf, & Cohen, 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.

3. Cognitive premises of the present research

Discourse in natural language incorporates cognitive mechanisms which, on the one hand, compel the speaker to choose metaphor when encoding the message, and on the other, they enable the recipient to decode the information transmitted in such a fashion and to tell metaphor from non-metaphor.

When we attempt to create a computational model of metaphor decoding process, we essentially make predictions about the inventory of cognitive factors which make processing of metaphor by humans possible. Computational experiments strip metaphor of the impacts imposed on it by the genre, the global context and the setting of the speech act, the implicit knowledge and presuppositions held by the speaker and the recipient about the topic – and thus allow us to draw inferences about the structure and the salience of signals of metaphoricity.

The computational experiment presented in this thesis looks at metaphor in the context of one sentence; we test the cues found in the broadly defined lexico-semantic and morphological environment of a metaphorically used word. The lexical and morphological cues are quantified with the index of statistical association with metaphoric contexts; the semantic cues are captured with a) distributional semantic distances and b) indexes of concreteness which are computed with the help of psycholinguistic data about lexical 'thingness' and with distributional semantic models. The contextual vectors disregard the linear sequence of cues in the sentence, since they are computed as a bag of cues (lexical, semantic, or morphological, respectively). Moreover, the values of all cues of one type are collapse into a single value by averaging over the contextual window.

As we compare the performance of computational models which are informed with different types of cues, we seek to establish the role of each in transmitting the signal of metaphoricity. The averaged bag-of-items representation 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.

4. Outline and contributions of the present research

The central goal of this thesis is to provide in-depth linguistic analysis of features which can be implemented in machine learning experiments for metaphor identification; we do not aim to evaluate efficiency of applicable algorithms. We will focus on the issues of feature engineering for metaphor identification and give linguistic interpretation to the efficiency of the performance of the selected features. Particular attention will be devoted to assessing the generalizability of each type of feature, i.e. the degree to which they are likely to serve as predictors of metaphoricity across different collections of data. The types of features to be explored in this research are:

- Semantic similarity;
- Lexical co-occurrence;
- Morphosyntactic co-occurrence;
- Concreteness indexes;
- Occurrences of flag words (lexical signals) and quotation marks.

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;

5. Structure of the thesis

The remainder of this thesis is divided into three parts.

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 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 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.

The thesis concludes with Chapter IV which provides in-depth analysis of the linguistic factors determining the performance of the models. For each of the experimental features, we identify the linguistic units which are most probable to carry the signal of metaphoricity and make predictions about their generalizability. Finally, the chapter presents the conclusions of the thesis and makes suggestions for future research in the area of computational identification of metaphor.

6. Notes on collaboration and publications

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.

The most representative parts of the work presented in this thesis was published at conferences in the field of computational and theoretical linguistics. The list of publications is included below. Three papers are devoted to computational metaphor identification and two to annotation of metaphor in corpus.

- •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.
- •Badryzlova, Y., & Lyashevskaya, O. (2017). Metaphor Shifts in Constructions: The Russian Metaphor Corpus. The 2017 AAAI Spring Symposium Series: Technical Reports, 127–130. Retrieved from http://www.aaai.org/ocs/index.php/SSS/SSS17/paper/view/15244
- •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.
- •Badryzlova, Y., & Panicheva, P. (2018). A Multi-feature Classifier for Verbal Metaphor Identification in Russian Texts. Artificial Intelligence and Natural Language, 23–34. St. Petersburg: Springer.
- •Panicheva, P., & Badryzlova, Y. (2017). Distributional semantic features in Russian verbal metaphor identification. Computational Linguistics and Intellectual Technologies, 1, 179–190. Moscow.

Chapter II. Experimental corpus

1. Corpus design

1.1. Selection of data

The Russian corpus of verbal metaphor which was used for the experiments presented in this paper is similar in design to the TropeFinder (TroFi) dataset of Birke and Sarkar (2006-- see Section ... of Chapter I). Since the corpus was compiled for the experiments in sentence-level identification of linguistic metaphor with a multiple-feature classifier described in the next chapter of the present thesis, the unit of the corpus is a single standalone sentence – as provided by the ruTenTen11 corpus (see below) from which the sentences were selected; sometime they correspond to lengthier stretches of text, when the context presented a problematic case for the parser which was used to preprocess the corpus – such as Sentences [3][4] below. The corpus contains annotation of only linguistic verbal metaphors at the sentence level: we look at the target verb in the context of one sentence and decide whether it is used metaphorically or non-metaphorically; the entire sentence is labelled as either 'MET' or 'NONMET' correspondingly.

The sentences for the corpus were selected manually by one expert (the author of this thesis) from the 14.5bn web-crawled corpus ruTenTen11 accessed via the SketchEngine interface (Kilgarriff et al., 2014). The expert analyzed the output of SketchEngine to the target verb query sentence by sentence and tagged each sentence as either metaphoric ('MET') or non-metaphoric ('NONMET') (for description of the metaphoric and non-metaphoric classes, see Section 2 of the present chapter). When reaching an equal number of MET and NONMET sentences, the expert stopped adding new sentences. RuTenTen11 contains no meta-data about the texts, so the genre or domain affiliation of the selected sentences cannot be precisely defined; however, in our judgement, they belong to a wide spectrum of genres and topics, such as web forums/social media, newspaper/journal articles, fiction, technical and academic texts, and others.

Not infrequently, the sentences contain typos, language errors at various levels, irregular usage (such as slang, jargonisms and colloquial style, absence of punctuation or incorrect punctuation, etc.). A remark on the notation of the examples: throughout the text of the thesis we will be using angular brackets ('<…>') to indicate the target verb in the sentence. Consider the following instances of irregular usage from the experimental corpus e.g.:

[1] (MET*) Решил рассказать о том, как здорово можно < жонглировать > рабочими столами имея мышку с доп и композом на бортовом компьютере.* (an incomprehensible word and a professional jargonism).

[2] (MET) *Уже полчаса < утюжу > место разными воблерами*. (A low-frequency narrow-domain term).

[3] (NONMET) *Моего сына как-то < укололи > когда нужно было температуру сбить мучали три дня на ушах стояли не спали вооообще даже смотреть не стала на это убежала в другую комнату муж был с максом сыниного плача на стены полезла остановить врачей никак уже почти 40 был на градуснике не помогали ни один мучался с зубами меня поймут видела его заплаканные глаза с слезах дикий крик венки набухшие на голове из-за крика. и когда маманьки без острой надобности начинают детей колоть меня бесит и хочется сказать - идите < уколитесь > сами если вам так не в терпеж... все простите за ругать* (Absence of capitalization (uppercase letters), missing punctuation at the boundaries of clauses, which seriously hinders the understanding, spelling mistakes);

[4] (NONMET) *мы были в больничке из за температуры ,,она снизилась только в понед после укола анальгина с супрастном Скорая повезла сначала в филевскую больницу ,там нас не приняли повезли на Сокол в икб1Приезжала та доктор .которая у нас была сДусей ,уже 2 раза узнала нас ,выругалась на нашего участкового -лечить не хочет ! в больнице Марусе < укололи > антибиотик цифазолин и ее обсыпало всю отменили .давали супрастин ,бифидум .а сегодня я под расписку домой ушла ,мы лежали во взр отделении ,правда, в отдельной вип палате на 2 человека бесплатно Дусю водили к лору .сказала надо вырезать аденоиды(((((((((((плохо дышит ..не знаю с садиком быть ее оттуда или пусть день в неделю ходит ?* (Inconsistent capitalization, erratic punctuation and absence of marked clause boundaries, truncated words, unconventional spelling of acronyms, spelling mistakes).

The motivation for sampling sentences from a web corpus and for including non-normative contexts into the experimental corpus stemmed from the intended purpose of the classifier – to identify metaphoric occurrences in raw running text, i.e. in the setting that approximates the real-life performance of an NLP application.

The total size of the corpus is 7,166 sentences; the corpus is balanced by the class (50% of the sentences are metaphoric and the other 50% are non-metaphoric). At the same time, the corpus is not balanced across target verbs, with the number of sentences per verb ranging between 225 and 693. However, this imbalance has been shown not to cause bias in the classification results: in an auxiliary experiment, the classifier was run on a randomized sample of the main corpus, which was balanced both across target verbs (224 sentences per verb) and for the MET and NONMET classes (112/112 sentences of each class per each target verb). The results obtained in this experiment proved consistent with the performance of the classifier on the unbalanced corpus – see Section … of Chapter III.

The experimental dataset is freely available for download and can be accessed at the following link: https://github.com/yubadryzlova/metaphor_dataset_20_verbs.git

1.2. Selection of target verbs

The corpus is built around 20 polysemous transitive Russian verbs. We aimed to select verbs with strong metaphoric potential, i.e. the verbs in which their metaphoric and non-metaphoric meanings are distinctly juxtaposed. The selected target verbs can be characterized with the following set of linguistic features:

1. the verb has at least one primary meaning which is a typical meaning of Accomplishment or Activity (Mehlig, 1985; Vendler, 1957);

2. the verb has at least one primary meaning which authorises a two-actant construction with the following mandatory arguments: (1) the Agent, (2) the Patient / the Theme;

3. the Agent of the primary meaning denotes a human being(s); the other arguments refer to physical (concrete, non-abstract) entities;

4. the derivational structure of the verb's polysemy is transparent: each secondary meaning is derived from the primary one by means of either a metaphoric or a distant metonymic shift;

5. the verb has a small number (< 10) of meanings listed in the Dictionary of the Russian Language[1] (Yevgenyeva, 1981);

6. the verb does not possess any strongly delexicalized meanings.

The meaning(s) of a target verbs that corresponds to criteria 2-3 above will be referred to as the **central literal meaning** in the subsequent sections. As might be seen, our definition of the central literal meaning overlaps with the notion of basic meaning in MIPVU (see Section …of Chapter I) with regard to the presence of the obligatory physical / concrete / human-oriented semantic component: as the MIPVU definition of the basic meaning has it: " … the basic meaning is a more concrete, specific, and human-oriented sense in contemporary language use" (G. Steen, Herrmann, Kaal, Krennmayr, & Pasma, 2010).

[1] This dictionary is conventionally known as the Minor Academic Dictionary whose title is abbreviated to MAS in Russian. We will use this acronym in the subsequent references to this dictionary in the rest of the paper.

The frequency of the selected target verbs in the range from 13.7 ipm (*vzorvat*) to 0.5 ipm (*bombardirovat*) according to the data of the Russian National Corpus (Lyashevskaya & Sharoff, 2009).

In the case of the verbs that have both the perfective and the imperfective aspectual forms, occurrences of both forms were included into the corpus: *napadat*: нападать / напасть, *ochertit*: очерчивать / очерить, *podkhvatyvat*: подхватывать / подхватить, *raspylyat*: распылять / распылить, *razbavlyat*: разбавлять / разбавить, *syedat*: съедать / съесть, *vykraivat*: выкраивать / выкроить, *vzorvat*: взрывать / взорвать, *vzvesit*: взвешивать / взвесить, and *zazhigat*: зажигать / зажечь.

The frequencies of different senses of polysemous words have different distributions in natural discourse (Lopukhina & Lopukhin, 2017; Lopukhina, Lopukhin, & Nosyrev, 2018). According the Sense frequencies project ('Sense frequecies with Russian Active Dictionary', n.d.), the four meanings of the verb *bombardirovat* 'to bombard smb / smth' are distributed in the ruTenTen11 corpus with the following ratio: 0.48 – 0.28 – 0.12 – 0.12; the two meanings of the verb doit 'to milk' are distributed as 0.71 – 0.29. When selecting sentences for our experimental corpus, we did not control the ratio of meanings in order to maintain the consistency with the goal of the experiment – to imitate the work of the machine learner in real-data conditions, i.e. when operating on raw running text.

The list of target verbs and the number of sentences in the subcorpus of each verb is available 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
очертить	ochertit	to outline (smth)	225
отрубить	otrubit	e	377
пилить	pilit	to saw (smth)	310
подхватывать	podkhvatyvat	to catch (smth falling)	373
причесать	prichesat	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
уколоть	ukolot	to prick (smth/smb)	353
утюжить	utyuzhit	to iron (clothes)	364
выкраивать	vykraivat	to cut (in sewing: parts of a garment, from fabric)	253
взорвать	vzorvat	to blow (smth) up, to explode (smth)	289
взвесить	vzvesit	to weigh (smth)	298
зажигать	zazhigat	to ignite (smth)	294
жонглировать	zhonglirovat	to juggle (smth)	396
Total:			7,166

2. Corpus annotation

The entire corpus was annotated by one expert; since the metaphor identification classifier was conceived as a binary classifier, the corpus annotation was carried out in terms of the two classes – metaphoric (‘MET’) and non-metaphoric (‘NONMET’). Thus, the annotator was compelled to make binary decisions about the metaphoricity of the verbal usages (i.e. either ‘MET’ or ‘NONMET’). In dubious cases, the decision still had to be made in favor of one of the choices.

The decision about assigning a sentence to either of the classes was made following the guidelines of the MIPVU procedure (G. Steen et al., 2010, p. 33, see Section … of Chapter I for more detail) which were slightly modified considering the fact that all the target verbs in our corpus were guaranteed to have a central literal meaning; this allowed us to omit the second step of the origingal MIPVU protocol:

1. Identify the contextual meaning of a lexical unit;

2. Determine whether the central literal meaning is sufficiently distinct from the contextual meaning;
3. Examine whether the contextual meaning can be related to the more basic meaning by some form of similarity.

2.1. Non-metaphoric class

The Non-Metaphoric Class was assigned sentences where:

1. the target verb was used in the central literal meaning (as described in Section 1.2), e.g.:

   [5] (NONMET) *После того , как вы уже < очертили > карандашом контур , слегка припудрите губки…*

2. the target verb was used 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, e.g.:

   [6] (NONMET) *Для этого лучше использовать угольный карандаш . Он четко < очертит > мелкие детали и придаст картине законченность и филигранность .*

   [7] (NONMET) *Рассматривая очертания распространяющегося по воздуху дыма , Денис Гур < очертил > рукой невидимый контур .*

2.2. Metaphoric class

The Metaphoric Class was assigned the following three types of sentences:

1. Conventionalized metaphors;
2. Unconventional creative metaphors;
3. Idiomatic expressions.

2.2.1 Conventionalized metaphors

Conventionalized Metaphors are target verbs used in one of their secondary meanings which is attested in dictionaries. In terms of the classification by Gurin and Belikova (2012) presented in Section …of Chapter I, this subclass includes latent metaphors. Recall that latent metaphors are defined as the meanings which are attested in dictionaries along with their literal meanings; a latent metaphoric word is not the only term with which a given referent can be denoted: in the context it can be replaced with its non-metaphoric counterpart without loss of the denotational meaning. For example, consider the latent metaphoric meanings of *trubit'* 'to blow a trumpet':

- to talk profusely about smb, smth; to spread gossip, information, news, etc.[2]; e.g.:

  [8] (MET) *СМИ < трубят > о достижениях в решении различных социально-экономических проблем.*

- to perform a tiresome or tedious action during a long period of time, e.g.:

  [9] (MET) *Так что, если даже … и получу по максимуму, останется < трубить > всего пять месяцев.*

2.2.2. Unconventional creative metaphors

Unconventional Metaphors exploit the verbs creatively to liken concepts from the Target domain to concepts from the Source domain and to reinterpret the Target in terms of the Source. Here we would like to make an important disclaimer concerning the methodological foundations of the present research: since our project aims to explore linguistic metaphor and its capacity to be identified by machine learning methods, we make no presumption as to the presence or absence of conceptual metaphoric mappings behind the examined linguistic metaphors. Therefore, the terms 'conceptual metaphor', 'Source' and 'Target are used in the subsequent paragraphs merely for the purposes of demonstration – in order to highlight the differences in the lexico-semantic complexity of metaphoric sentences.

Unconventional metaphors vary in the degree of their unconventionality (creativeness); Gurin and Belikova (2012) distinguish rare, novel, innovative, and creative metaphors.

As we discussed in the previous chapter (Section …), **rare metaphors** are defined as metaphoric expressions which are repeatedly (although not highly frequently) attested in the corpus. These are non-unique, genre-specific metaphors which occur in texts over a certain period of time. As a rule, they represent an expected development of the conceptual metaphors that already exist in a language (Gurin & Belikova, 2012). We suggest extending the scope of this definition as follows: rare metaphors project one of the facets of the Source domain which has already been exploited in more conventional metaphors to a new Target domain. For instance, consider the following examples of rare metaphors with the verb *pilit* 'to saw (smth)':

[2] MAS, the Dictionary of the Russian Language (Yevgenyeva, 1981)

[10] (MET) *Останутся одни " эффективные менеджеры " , только и умеющие " < пилить > деньги " и брать откаты .* ‘There will remain only ‘effective managers’ which know only one thing – how to < saw > (lit. embezzle) funds and to take kickbacks.’

Here, the rare metaphor is based on the facet of the central literal meaning which profiles sawing as an action aimed at separating the whole into several parts; the new Target domain is VALUABLE MATERIAL RESOURCES (predominantly, money and property).

Another set of rare metaphors triggered by the verb *pilit* exploits the facet of the Source domain which profiles the action of sawing as a prolonged, effort-intensive, and energetic physical activity, e.g.:

[11] (MET) *Сам < пилит > сервера (ставит , настраивает , оптимизирует).* ‘He < saws > (lit. services) servers himself (installs them, configures, and optimizes).’

[12] (MET) *… < пилим > вики - разметку , вставляем ссылки на редиректы – готово!* ‘We < saw > (lit. develop) Wiki-markup, insert links to redirect pages – that’s it!’

The new Target domain here can be loosely defined as HIGH-TECH SERVICES gererated as a result of an activity.

Novel metaphors are defined as metaphoric expressions which belong to the most recent period of time and whose frequency in the corpus is increasing over the time. They can either express conceptual metaphors that have already been attested in a language, or they can be manifestations of new emerging conceptual mappings.

Innovative metaphors are lexically unique, very infrequent, but conceptually predictable metaphoric expressions which emerge when a slot of the Target domain is filled with a non-trivial word representing the Source domain; thus, innovative metaphors are metaphors where lexical lacunas in the conceptual metaphors are filled idiosyncratically (Gurin & Belikova, 2012). Consider the examples of novel and innovative metaphors (Sentences [13]and [14]):

[13] (MET) *Сестра поглядела на нее, словно < уколола > кинжалом.* ‘Sister threw a glance at her, as if she < pricked > her with a dagger.’.

This sentence contains a realization of the conceptual metaphor GLANCE IS A SHARP INSTRUMENT where the frame ‘Instrument’ is reinforced with one more word from the Target domain of sharp instruments – ‘dagger’ (compare sentence [13] to its less conceptually complex equivalent *Сестра уколола ее взглядом* ‘Sister pricked her with a glance’). The sentence under examination also contains a metaphor flag – the word *словно* (‘as if’). Notice that both the Source (‘glance’) and the Target (‘dagger’) domains are equally represented, each by one lexeme.

[14] (MET) *Что это должно было означать, я не совсем понял, но какая-то заноза всё-таки < уколола > сердце.* 'I did not quite understand what it was supposed to mean, but still I felt a splinter < prick > my heart.'

This context exemplifies the established conceptual metaphor STRONG EMOTIONS ARE SHARP OBJECTS; notice that the Target domain (EMOTIONS) is not manifested on the lexical level: the Target is constructed by the entire thematic framework of the sentence. Meanwhile, the Source domain is reinforced by a lexeme representing its 'Instrument' frame ('splinter'). Compare sentence [14] to its less conceptually complex paraphrase:

[15] (MET) *Подозрение укололо мое сердце* . 'Suspicion has pricked my heart'

– which contains only words from the Target domain.

For more examples demonstrating novel and innovative metaphors, see also sentences [22]-[23] which are discussed below.

Creative metaphors are unique both conceptually and lexically; the Source – Target pairs connected by their conceptual mappings are not rooted (or very weakly rooted) in the conceptual system of the language, e.g.:

[16] (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.'.

This is a highly creative and conceptually complex metaphor which contains more than one nested mappings and their entailments. The primary mapping VANITY IS A BALOON is grounded in the notion of balloon as being inflated, empty of content, lightweight and vulnerable. The surrealistic transformation of this Source which reconceptualizes the balloon as a container for wind – adds an extra layer to the metaphor by building up a new frame on top of the Source concept. The entailments of the reframing include such concepts as unpredictability, increase in size, and potential destructiveness. Finally, all these ingredients resolve into the following succession of mappings: HURTING ONE'S VANITY IS PRICKING A BALOON and HURT VANITY IS A STORM.

[17] (MET) *Игорь Забегин сравнивает сети предприятия с человеческим организмом: < уколол > палец, а почувствовал всем телом.* '[Mr.] Zabegin compares a business's network to a human body: if you < prick > you finger, your entire body senses [the pain].'

This sentence represents a surface realization of the conceptual metaphor ORGANIZATIONS ARE LIVING ORGANISMS; both the Source and the Target are explicitly expressed, although the Target ('a business's network') is underrepresented with respect to the Source ('human body' ... 'body', 'finger', 'pain'); the operator of comparison *сравнивать* 'to compare' smth to smth) underscores the analogy between the Source and the Target.

2.2.3. Idiomatic Expressions

The third type of sentences tagged as belonging to the Metaphoric Class (along with conventionalized and unconventional metaphors presented above) are idioms, i.e. fixed or semi-fixed non-compositional units. We regard idioms as a subclass of metaphoric expressions following Lakoff and Johnson who underscore the metaphoric nature of idioms, since their meaning is motivated by metaphorical mappings and certain conventional mental images: "… the relationship between the meaning of the parts and the meaning of the whole is complex. The words evoke an image; the image comes with knowledge; conventional metaphors map appropriate parts of that knowledge onto the target domain; the result is the meaning of the idiom. Thus, a metaphorical idiom is not just a linguistic expression of a metaphorical mapping. It is the linguistic expression of an image plus knowledge about the image plus one or more metaphorical mappings" (Lakoff & Johnson, 1999, pp. 69–71).

Besides, our decision to include idioms into the experimental corpus was guided by goal of the experiment – to identify metaphoric occurrences in raw text.

Idiomatic expressions often inherit the constructional properties (the models of syntactic and lexical combinability) from the primary meanings of their constituent words, which poses an additional challenge to the task of automatic metaphor classification. For example, the construction of the Russian idiomatic expression *греть руки на чем-л.* 'to warm one's hands with smth' (meaning 'to make dishonest or illegal profit') is identical to the construction of its non-metaphoric counterpart; the two constructions differ only in the semantic and lexical class of the indirect object – cf. Sentences [18]and [19]:

[18] (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.'

[19] (NONMET) *... мы ... стояли у окна ... и < грели > руки на батарее .* We were standing by the window warming our hands by the radiator.

This subclass of our metaphoric class includes idiomatic expressions that are attested in dictionaries, as well as imagery-based collocations which are not attested in lexicographic sources; however, the latter collocations occur in our experimental corpus or in the Russian National Corpus on a regular basis, and therefore we tag them as idiomatic subclass; consider, for example, the expression *взорвать (кому-л.) мозг / мозги* ‘to explode (one’s) brain’; lit. ‘to astound smb’, cf. *вынести кому-л. мозг* lit. ‘to drive one’s brain out of one’s head’, e.g.:

[20] (MET) *«Эта книга < взорвет > вам мозг», — примерно так любят писать в аннотациях к большинству современных книг, выходящих в России.* “This book will blow up your brain “ – these words have become a catchphrase for the blurbs written for most of contemporary fiction books currently published in Russia.’

Sometimes the imagery contained in an idiom may give rise to a metaphor which becomes an extension of the idiom:

[21] (MET) *Слово может так твой мозг < взорвать > , что все вокруг забрызгает.* ‘Word is capable of < exploding > your brain (lit. astounding you) so violently that it will splatter everything around.’

[22] (MET) *Чтобы выдержать должную последовательность в своей интерпретации аграрной проблематики ..., марксизм был вынужден < причесать > под одну стальную гребенку остальные сюжеты, так или иначе связанные с деревней.* ‘In order to be consistent in its interpretation of the agricultural agenda, Marxism had to use a single comb of steel in order to < comb > (lit. to equalize, with a negative connotation) all the remaining issues which were somehow related to rural affairs.’

Sentence [21] contains the idiom *взорвать (кому-л.) мозг / мозги* lit. ‘to explode (one’s) brain’ discussed in the previous paragraph; on top of it, there is a metaphor which exploits the Target domain of the original idiom: if a brain explodes, the splashes of it are likely to cover the surrounding objects.

In sentence [22] the metaphor of the comb made of steel – which is used to express the Source domain of a ruthless policy – is triggered by the idiom *причесать всех под одну гребенку* ‘to comb everyone with the same comb’ lit.‘to make everybody equal, disregarding their individual features’.

2.3. Distribution of metaphoric subclasses in the corpus

Distribution of metaphoric subclasses in the corpus is shown in Table 2.

Table 2. Distribution of metaphoric subclasses in the metaphoric subcorpus

Russian	transliteration	conventionalized metaphors, % of MET sentences	unconventional metaphors, % of MET sentences	idiomatic expressions, % of MET sentences
бомбардировать	bombardirovat	94.43	5.57	0
доить	doit	78.59	20.13	1.28
греть	gret	83.7	11.53	4.77
нападать	napadat	78.27	6.39	15.34
очертить	ochertit	95.56	0	4.44
отрубить	otrubit	92.04	6.9	1.06
пилить	pilit	98.06	1.29	0.65
подхватывать	podkhvatyvat	83.92	8.04	8.04
причесать	prichesat	68	6	26
распылять	raspylyat	97.19	2.81	0
разбавлять	razbavlyat	100	0	0
съедать	syedat	80.67	11.83	7.5
трубить	trubit	84.38	15.62	0
уколоть	ukolot	52.41	47.59	0
утюжить	utyuzhit	96.15	3.3	0.55
выкраивать	vykraivat	99.21	0.79	0
взорвать	vzorvat	89.62	2.08	8.3
взвесить	vzvesit	71.14	0	28.86
зажигать	zazhigat	80.95	19.05	0
жонглировать	zhonglirovat	95.45	4.55	0
Total:		92.5	4.8	2.7

Not unexpectedly, conventionalized metaphors by far outnumber the two other classes – both in the datasets of individual target verbs and in the entire corpus. However, the ratio of unconventional metaphors and idioms substantially varies across the verbs – within the ranges of 0 – 48% and 0 – 29%, correspondingly. This is presumably a benign factor for the generalizability of machine learning models trained on the corpus, since it ensures diversity in the distribution of different types of metaphoric occurrences.

The verb *vzvesit* ‘to weigh (smth)’ has the highest percentage of idioms (28.9%) due to the high frequency of the idiomatic expressions *взвесить все за и против / все плюсы и минусы* ‘to weigh all the pros and cons / all the pluses and minuses’; lit.‘evaluate a situation or circumstances by means of carefully considering the advantages and disadvantages of smth’. It is followed by the verb *prichesat* ‘to comb (smth/smb)’ with 26% of idiomatic expressions comprised predominantly by the phrase *причесать всех под одну гребенку* ‘to comb everyone with the same comb’; lit. ‘to equalize everybody, disregarding their individual features’ (cf. *стричь всех под одну гребенку* ‘to cut everyone’s hair using the same comb’). The third verb with a high ratio of idioms (15.4%)

is *napadat* ‘to attack (smth/smb)’ with its phraseological expressions a) *напасть на (чей-л.) след (напасть на след кого-л.*) ‘to pick up smb’s / smth’s trail’; lit. ‘to discover the signs of smb’s / smth’s presence, influence, etc.’, b) *не на того напал(-а, -ли)!* ‘the person you are attacking is not what you expected to be’; lit. ‘the person you are dealing with is different from your expectation of him/her; you underestimate the person you are dealing with’, and c) жор напал ‘(insatiable) hunger has attacked (smb)’; lit. ‘smb couldn’t stop eating’. In total, about 3% of sentences in the corpus contain idiomatic expressions.

The highest amount of unconventional metaphors is contained by the subcorpus of the verb *ukolot* ‘to prick (smth/smb)’; this verb is frequently used in novel metaphors which evoke multiple frames of the Source domain (such as the pain caused by pricking, the parts of body exposed to pricking, the sharp instrument used for pricking, etc.) – see, for example, sentence [14], or:

[23] (MET) *Его шутка < уколола > меня в самое больное место.* His joke < pricked > me in my sorest spot.

There is a number of cases of extended metaphor, when the Source domain (PRICKING INSTRUMENT) attracts frames from other Source domains; consider Sentence [24] where the Target domain concept *ревность* ‘jealousy’ participates in three mappings at once: it is framed as a pricking instrument, as a burning substance (or maybe object), and as an iron hand::

[24] (MET) *Ревность < уколола > Сэмми прямо в сердце, потом она прожгла его до желудка и железной рукой сжала внутренности.* Jealousy < pricked > Sammy straight in the heart; then it burnt him all the way down to his stomach and clenched his guts with an iron hand.

Another verb with a high ratio of unconventional metaphoric sentences (20%) is *doit* ‘to milk (e.g. a cow)’. The most widespread conceptual mapping with this verb in the corpus is A PERSON / A GROUP OF PEOPLE (also metonymically, e.g. a nation, a country, a company, etc.) IS A DAIRY COW (which is mediated by another metaphor, A PERSON / A GROUP OF PEOPLE IS A GENERATOR OF PROFIT):

[25] (MET) *Нынешние власти ... призваны сохранять этот статус-кво, при котором страна находится в состоянии "дойной коровы", которую < доят > , покуда она доится ...* The mission of the current authority is to maintain the status quo which ensures that the country remains a “dairy cow” which can be < milked > as long as it yields milk…

The verb *zazhigat* ‘to ignite (smth)’ also has a large number of unconventional metaphoric sentences (19%); it behaves similarly to *ukolot* above: it evokes multiple frames of the Source domain (such as the flame or the light resulting from the action of igniting, proliferation of fire, etc.), e.g.:

[26] *(MET)... Были < зажжены > и до сих пор продолжают гореть, вспыхивая кровавыми сполохами, национально конфликты.* The ethnic conflicts that were ignited at that time are still burning belching blood-red flames.

There are three verbs for which no unconventional metaphors have been attested in our experimental corpus: *ochertit* ‘to outline (smth)’, *razbavlyat* ‘to dilute, to liquefy (smth)’, and *vzvesit* ‘to weigh (smth)’ – these verbs evoke only conventionalized metaphors.

In total, about 5% of all the metaphoric sentences in the corpus contain unconventional metaphoric expressions.

3. Annotation reliability test

3.1. Selection of sentences

As we mentioned above (Section 1.1), annotation of the entire corpus was performed by one annotator (the ‘primary annotator’). In order to test the quality of annotation, we implemented a reliability test in which two new analysts independently annotated 20% of sentences from each verb (a total of 1,425 sentences).

Prior to the reliability test, the primary annotator evaluated each sentence from the large, 7K sentence corpus, in terms of their difficulty for judgement of metaphoricity by human subjects. Sentences were tagged as ‘difficult’ if:

– A sentence is incomplete, contains omissions, ellipses, typos, errors, unconventional acronyms or proper names, and homonymous words that can not be disambiguated outside of broader context (e.g., see sentences [1]-[4]);
– A sentence is very short (4-5 words or less) and the target verb can not be understood outside of broader context, e.g. *Он уже все < взвесил >* . ‘He has already < weighed > (or lit. carefully considered) it all.’;
– One or more of the target verb’s actants is expressed metonymically, e.g. … Детская непосредственность и откровенность артиста в очередной раз < взорвала > зал. ‘The childlike straightforwardness and frankness of the singer once again < exploded > the hall (lit. provoked a strong emotional reaction of the audience).’

- The meaning in which the target verb occurs in the sentence is related to the central literal meaning by a close metonymic shift (e.g. < *Трубя* > *хоботом, слон подает различные сигналы* ... ‘By < trumpeting > with his trunk, an elephant emits different kinds of signals’.);
- The meaning in which the target verb occurs in the sentence is closely associated with the central literal meaning but it licenses a grammatical construction which is different from the two-actant construction of the central meaning, cf.: *Мы* < *греем* > *на сковородке масло* ... ‘We < heat > oil on the frying pan…’ (two-actant construction, central literal meaning) vs. *солнце не* < *грело* >*, а просто светило* . ‘… the sun was not < heating > (lit. the sun was not hot), it was only shining.’ (one-actant construction);
- If the conventionalized metaphoric meaning presumably has a narrow scope of usage (for example, they belong to domain-specific, e.g. professional, discourse, such as a jargonism, a technical term, etc.) – see, for example, sentences [11]-[12];
- The target verb was used in the sentence as a part of an idiom or an unconventional metaphor – see, for example, sentences [13]-[17], [21]-[22].

In total, 22.8% of the corpus were tagged as ‘difficult’. A randomized dataset was generated: approx. 20% of sentences from each verb were drawn so that ‘difficult’ metaphoric, ‘difficult’ non-metaphoric, ‘not difficult’ metaphoric and ‘not difficult’ non-metaphoric sentences were represented in equal proportions (25% from each of these subclasses). The resulting dataset contained 1,425 sentences which were offered to the two new annotators to tag.

3.2. Annotator instructions

The two new annotators were native speakers of Russian with initial training in linguistics (students who have completed their first year of undergraduate Linguistics program) but with no background in metaphor theory. They received no specialized training targeted at practicing and testing their familiarity with the annotation protocol prior to the experiment.

The annotators were working in the Google Spreadsheet environment; each row contained one sentence (where the target verb was indicated with pointy brackets and spaces) and a drop-down menu from which they chose the values to indicate the degree of metaphoricity of the target verb. The annotators received two pages of written instructions (the full version of the annotator’s guidelines is available in Appendix 1).

We split the annotation task into the two subtasks; the annotators were asked to:

1) Separate the metaphoric and non-metaphoric contexts;

2) Identify the degree of their confidence in the attributed class on the five-point scale (see below). This step was motivated by the desire to make the procedure more suitable for non-professional annotators, since the range of several options provides opportunity for tagging borderline and dubious cases.

A major change was introduced into the first step of the original annotation procedure. We inversed the metaphor identification task: we asked the annotators to tag **all the non-metaphoric, literal** occurrences of the target verbs. Thus, the annotators' actions under the inversed task formulation were supposed to follow the procedure which could be described as follows:

1. Identify the basic meaning of the lexical unit. The basic meaning is a more concrete, specific, and human-oriented sense in contemporary language use;
2. Identify the contextual meaning of the lexical unit;
3. Determine whether the basic meaning is sufficiently similar to the contextual meaning;

The inverted task has been largely made possible by the overall setup of the annotation experiment:

- the annotators looked at one sentence and one target verb at a time;
- the sentences were grouped by the target verbs; this enabled the annotators to maintain consistency in their decisions: they had to identify the basic meaning (step 1 of the inversed procedure) only once for each group, and then for each sentence of the group they needed to repeat only steps 2-3;
- the number of target verbs and, consequently, the groups of sentences, (20 in total) was relatively low, which decreases the number of unique decisions required from the annotators – as compared to annotation of running text.

Since the annotation guidelines were designed in order to be accessible to non-professional annotators, we chose an informal way of exposing them to the instructions, which relies on the intuitive concepts which are supposed to be understood by most language users.

To illustrate the difference between metaphoric and non-metaphoric sentences, we used the verb *состряпать* 'to have smth baked' (this verb appeared only in the guidelines and did not appear in the corpus). As an example of a non-metaphoric sentence, the annotators were exposed to the following context: *Мартин < состряпает > тыквенный пирог, а Хина приготовит сочный бифштекс.* 'Martin < will have baked > a pumpkin pie, and Khina will < have cooked > a juicy beefsteak'; this sentence was accompanied by a picture of a happy cook with a dish of pies. The annotators were asked to picture in their minds the most prototypical direct meaning of the target verb each time they made a decision. As an example of a metaphoric sentence, we presented the following: *Вместе они быстро < состряпали > текст письма*. ‘Together they were quick < to have baked > (*lit.* concocted) the text of the letter.’

The fourth step of the original MIPVU procedure (‘Examine whether the contextual meaning can be related to the more basic meaning by some form of analogy.’), which is intended to identify unconventional metaphors, was formulated in the annotator guidelines as follows: the annotators were instructed to tag as metaphorical any sentences containing unconventional highly creative metaphor as well as idiomatic expressions. Such form of instruction relied on the expectation that an average native speaker of the language with completed high school education should be familiar with the notions of idiomatic expression and creative metaphor.

In the second step of the annotation procedure (identifying the degree of confidence) the annotators were offered a choice of five categorical classes (as compared to the binary classes in the first round of annotation):

- ‘Direct Meaning’,
- ‘Indirect Meaning’,
- ‘Likely Direct Meaning’, ‘
- Likely Indirect Meaning’,
- ‘The Meaning of the Sentence is not Clear’ (the annotators were also asked to restrict the use of the latter tag to 2-4 times per each target verb).

3.3. Binarization of categorical annotation

The results produced by the two annotators were expressed in categorical terms (the choice of five options), therefore they had to be made compatible with the binary annotations of the principal annotator. We applied two **binarization schemes**.

The first scheme followed the straightforward logic: all the ‘Likely Direct Meaning’ tags were converted to ‘Direct Meaning’; all the ‘Likely Indirect Meaning’ tags were converted to ‘Indirect Meaning’. In cases when one of the annotators chose ‘The Meaning of the Sentence is not Clear’ we looked at the choice of the second annotator. Information about the tags from the referee (the principal annotator) was not taken in consideration, therefore, this binarization scheme was not conducive to the opinion of the referee.

The second, more sophisticated scheme was favorable towards the choices made by the referee: if one of the Annotators chose 'Likely Direct Meaning' and the other chose 'Likely Indirect Meaning', the scheme looked at the opinion of the referee to convert them to either 'Direct' or 'Indirect' tags. Similarly, in the cases of 'The Meaning of the Sentence is not Clear', we looked at the annotation produced by the referee.

3.4. Annotation results and analysis

3.4.1. Inter-annotator agreement

The inter-annotator agreement between the three annotators was measured as Fleiss κ (Siegal, 1956). When the two annotations from the new annotators were binarized according to the complex ('conducive') scheme the agreement amounted to 0.9, while the agreement obtained on the annotations binarized according to the straightforward ('strict') scheme equaled 0.83. Both results indicate strong degree of agreement.

The obtained result cannot be directly compared to other metaphor annotation projects due to differences in their design.

The series of annotation reliability tests with the MIPVU procedure on English and Dutch texts reported by Steen et al. (2010) yielded the Fleiss κ ranging between 0.79 and 0.88 on the English data, and between 0.83 and 0.92 on the Dutch data. Both the English and the Dutch tests were performed in the running-text paradigm, where the analysts had to identify all the metaphoric occurrences in fragments of texts whose length varied between approximately 900 and 1950 words. The English texts were selected from the four registers (Fiction, Academic, News, and Converstation) of BNC Baby, a subset of the British National Corpus; the Dutch texts were drawn from a corpus of spoken Dutch (Corpus Gesproken Nederlands) and from a electronic database of major national newspapers. The parallel annotations in all reliability tests were provided by four analysts all of which were professional linguists with a high degree of competence in metaphor theory.

Klebanov and Flor (2013) report the agreement (κ) of 0.58 between two non-professional annotators who annotated all metaphoric occurrences in 116 test-taker essays written in English, totaling 55,473 tokens.

In the annotation experiment by Shutova and Teufel (2010) three lay annotators reached the κ agreement of 0.64 on the task of annotating all the verbs which are used metaphorically in a stretch of text (the size of the text is not reported).

As follows from the descriptions, the major difference between the multiple annotator experiments reported in the literature and our annotation reliability experiment is that the former projects were carried out on longer stretches of running text where annotators were asked to identify all the metaphoric occurrences in each sentence. Apparently, this is a much more cognitively complex activity, since the annotators have to make a large number of unique decisions about words from different parts of speech possessing various degrees of metaphoric complexity – as opposed to the structured method of presenting the material for annotation implemented in our reliability tests. Secondly, our annotators were given certain leeway in their decision making (the five-point scale of confidence), which presumably relieved them of the pressure of choosing between binary options in dubious cases.

In terms of qualifications of the annotators our annotation project occupies an intermediate position, as one of the annotators was a professional linguist, and two were non-professional (nearly naïve) – unlike in the other four projects above, where all of the participants were either professionals or lay persons.

3.4.2. *Inter-annotator disagreement: analysis*

The inter-annotator disagreements which remained after the new annotators' decisions were binarized according to the straightforward ('strict') scheme are summed up in Table 3. The table indicates that the two annotators tend to disagree more with the referee than with each other.

Table 3. Matrix of Disagreements, number of sentences (A1 – Annotator 1, A2 – Annotator 2, Ref – Referee).

Total: 1412 sentences

	A2	**Ref**
A1	43	85
A2		68

We have analyzed the cases where the annotators agree with each other and disagree with the referee (i.e. the intersection of the sentence sets in the rightmost column of Table 3) – 55 sentences in total.

The majority of the cases of disagreement occur on the two types of sentences:

- Sentences with conventionalized metaphors where the target verb is used in a borderline meaning, i.e. the meaning that possesses concrete semantics but lacks important semantic components that are present in the central literal meaning (for the working definition of the central literal meaning, see p. 11). For example: *канат < пилит > плечо* ‘the rope < saws > (lit. chafes) the shoulder’, *электроны < бомбарадируют > Землю* ‘electrons < bombard > the earth’, *мотор / самовар < трубит >* ‘the engine / the kettle < trumpets > (lit. produce a tubular sound), *лось < трубит >* ‘the moose < trumpets > (lit. bellows)’, *чайки < трубят >*‘*seagulls* < trumpet > (lit. squawk)’, *< утюжить > волосы* ‘to iron one’s hair (lit. to straighten one’s hair using a hair iron’).
- Sentences with novel, innovative, and creative metaphors, especially those that contain operators of comparison (‘flag words’), e.g.:

[27] *Один человек заявил, что когда у него нет Интернета, то возникает такое чувство, будто бы ему < отрубили > руку.* ‘One person said that when he has no access to the Internet, he starts feeling as if he had his hand < chopped off >.’

[28] *С тобой спорить — все равно, что дохлую корову < доить >.* ‘Arguing with you makes as much sense as < milking > a dead cow.’

[29] *Быть матерью, женой и бизнесвумен -это все равно что < жонглировать > несколькими предметами, не будучи к этому как следует подготовленной.* ‘Being a mom, a wife, and a business woman is similar to < juggling > several objects at once, without having been trained for it.’

The fact that the annotators chose to class unconventional metaphors of this type as non-metaphoric sentences can be accounted for by the two reasons. Firstly, the annotators were primed using non-metaphoric stimuli (see the annotation instructions, Section 3.2, and their task required them to identify all the non-metaphoric occurrences in the corpus. Secondly, the annotator guidelines gave no explicit instructions about such cases and provided no explicit examples of handling such sentences (recall that the only mention of creative metaphors in the guidelines invited the annotators to rely on their existing understanding of this term). Since metaphors of these types (especially when they are reinforced by flag words) are based on explicit analogy between Source and Target domains, and the target verb equally belongs to both, it is quite natural that inexperienced annotators tend to assign them to the non-metaphoric class.

Another possible reason behind the bias exhibited by the two new annotators is that non-professionals are more likely to traverse the text in a rapid skimming manner so that their eye fixations focus on the minimum context surrounding the target verb, i.e. for example, in case of

Sentences [27] and [28] they would capture the words belonging to the Source domain: < *отрубили* > *руку* ‘he had his hand < chopped off >’, and *дохлую корову* < *доить* > ‘< milking > a dead cow’. These non-metaphoric cues then presumably promote the annotator’s decision in favor of the non-metaphoric class.

Overall, the fact that the inter-annotator disagreements are localized within one particular subclass of metaphoric occurrences and the high values of Fleiss κ confirm that the annotation of the entire corpus may be considered reliable.

Summary of Chapter II

- We collected and annotated a corpus of approximately 7,000 Russian sentences each of which contains one of the 20 target verbs. The corpus is made publicly available.
- The target verbs are polysemous transitive verbs with strong metaphoric potential, i.e. the verbs in which their metaphoric and non-metaphoric meanings are distinctly juxtaposed.
- The subcorpus of each target verb is balanced by class: 50% of the sentences are metaphoric, and 50% are non-metaphoric. The metaphoric class is comprised by the three subclasses: conventionalized and unconventional metaphors, and idiomatic expressions.
- The test of the reliability of annotation in which three annotators independently analyzed 20% of the corpus yielded a high measure of inter-annotator agreement (Fleiss κ): 0.83-0.9.

Chapter III. Automated metaphor identification experiment

This chapter describes the main practical contribution of the present thesis – the experiment in identification of verbal metaphor in the annotated Russian dataset which was presented in Chapter II . The experimental task was defined as binary classification of sentences into those containing a metaphoric occurrence of the target verb vs. sentences with non-metaphoric occurrences of target verbs. We used Support Vector Machine (SVM) classifier with linear kernel and 5-fold cross-validation. The set of features was comprised of five types of features. The rest of the chapter is organized as follows: Section 1 presents the theoretical models which guided our choice of features which were implemented in the experimental classifier. Section 2 will give a brief outline of the techniques which were used to preprocess the experimental corpus. Section 3 contains a detailed account of the set of features which were engineered for the experiment; this part of the thesis will describe the computational models and the statistical measures deployed in order to obtain the vectors for machine learning. Section 4 briefly reports the technical details of the experimental setup. Section 5 offers a report of the results of the experiment and an in-depth discussion of these results, including efficiency and inefficiency of features and their configurations, a comparison of one-feature and multi-feature models, and an evaluation of the tradeoff between model complexity and the gain if efficiency.

1. Motivation behind the choice of features

The feature set which was used in our metaphor identification experiment consisted of five types of features:

– Distributional semantic feature
– Lexical co-occurrence feature
– Morphosyntactic co-occurrence feature
– Concreteness feature
– Flag words (lexical markers of metaphoric language) and quotation marks features

1.1. Motivation behind the use of distributional semantic feature

The reasoning behind application of distributional semantic features relies on several linguistic theories which can be divided into two major groups with regard to whether they focus on the semantic incongruity or the semantic cohesion of metaphoric contexts.

The theories of the first group treat metaphor as a form of contextual semantic anomaly – such are the componential models of word meaning and the related view on metaphor as a violation of selectional preferences (Katz & Fodor, 1963; Wilks, 1978), and the theory of conceptual metaphor (Lakoff & Johnson, 1980a, 1980b). The theories of the second group stress the contextual cohesion which arises from the semantic differences among the meanings of polysemous words – such as the models of semantic combinability suggested by the Moscow Semantic School (Apresyan, 1995, 2009), along with the concept of associative chains (Halliday & Hasan, 1976).

Componential models (Katz & Fodor, 1963) represent word meaning as a set of semantic markers and distinguishers. For example, the meanings of the word *bachelor* can be captured by the following sets of semantic descriptors:

1) (Human) → (Male) → [who has never been married];
2) (Human) → (Male) → (Young) → [knight serving under the standard of another knight];
3) (Human) → [who has the first or lowest academic degree];
4) (Animal) → (Male) → (Young) → [fur seal when without a mate during the breeding time] (Katz & Fodor, 1963, p. 190).

Apart from differentiating the meanings of polysemous words, the marker models of meaning are intended to explain the incongruity of otherwise well-formed sentences such as:

[30] The paint is silent. (Katz & Fodor, 1963, p. 175)

Semantic congruity is enforced in language by having words impose marker-based selectional constraints on each other as directed by the underlying syntactic structure. Thus, *silent paint* is considered aberrant because *silent* imposes the selectional constraint '+capable of producing sounds' on its head noun.

Basing on Katz and Fodor's approach to semantics, Wilks (1978) suggests the view of metaphor as a violation of a word's selectional preferences, where selectional preferences are the soft semantic constraints that a predicate places onto its arguments. Metaphor is viewed as a semantic anomaly, since metaphoric expressions are semantically deviant from their linear context. Consider the following frequently cited example:

[31] My car *drinks* gasoline. (Wilks, 1978, p. 199)

The verb *drink* normally requires a subject that is an animate being and a direct object that is a potable liquid. Therefore, *drink* taking a *car* as a subject is an anomaly which, according to Wilks, suggests the metaphoric usage of *drink* in this context.

The Conceptual Theory of Metaphor (Lakoff & Johnson, 1980a) metaphors arise when nonmetaphorical concepts are mapped onto metaphorical domains. Nonmetaphorical concepts (or Source domains) are “those that emerge directly from our experience and are defined in their own terms.” Some of nonmetaphorical domains are:

1) Spatial orientation (e.g. UP-DOWN, IN-OUT, NEAR-FAR, FRONT-BACK);
2) Ontological concepts arising in physical experience (e.g. ENTITY, SUBSTANCE, CONTAINER, PERSON);
3) Structured experiences and activities (e.g. EATING, MOVING, TRANSFERRING OBJECTS FROM PLACE TO PLACE, etc.).

Metaphorical concepts (or Target domains) are “those which are understood and structured not merely on their own terms, but rather in terms of other concepts. This involves conceptualizing one kind of object or experience in terms of a different kind of object or experience.” Examples of metaphorical concepts are: CONTROL, IDEAS, MIND, LIFE, TIME, etc. (Lakoff & Johnson, 1980b).

In the tradition of the integrative approach to the description of lexical systems which was developed by the Moscow Semantic School (see, e.g., Apresyan, 1995, 2009), predicates are analyzed in terms of valences and arguments. Each predicate possesses a number of semantic valences (their number ranging between one and six (Apresyan, 1995, pp. 135–137)); some of the types of semantic valences are: Subject, Object, Counteragent, Addressee, Recipient, Location, Itinerary, Instrument, Motive, Cause, etc. (Apresyan, 1995, pp. 125–126). Each valence can be filled only by certain words. The Moscow Semantic School delineates two types of combinability:

a) Lexical combinability, when a valence can be filled with words that do not share any common semantic components, e.g. the figurative meaning of the verb *сбрасывать* ‘to reduce smth; lit. to throw smth down’ can take only the following few words as its arguments: *давление* ‘pressure’, *газ* ‘gear’, *скорость* ‘speed’, *температура* ‘temperature’, and *вес* ‘weight’ (Apresyan, 1995, p. 61). This type of combinability apparently cannot contribute much to the use of distributional semantic features in our experiment; however, it may prove efficient when applying lexical unigrams as classification features – their implementation will be discussed below in Section 1.2.

b) Semantic combinability, when a valence the potential fillers of a valence share a common semantic feature, e.g. the verbs *улучшаться* ‘to improve (intransitive)’ and *ухудшаться* ‘to worsen, to deteriorate’ can apply only to arguments denoting conditions, capacities, and processes; they cannot combine with nouns of physical objects and persons: compare the well-formed Russian sentences [32] and [33] with sentences [34] and [35] which sound anomalous in Russian:

[32] *Погода ухудшилась / улучшилась*. ‘The weather got words / improved.’

[33] *Зрение / поведение ухудшилось / улучшилось*. ‘The eyesight / behavior deteriorated / improved.’

[34] **Ручка ухудшилась / улучшилась*. ‘*The pen got worse / improved.’

[35] **Петр ухудшился / улучшился*. ‘*Peter deteriorated / improved.’ (Apresyan, 1995, pp. 61–62)

Polysemy arises when one valence of a predicate can be filled with two or more distinctly different semantic classes. For instance, the non-metaphoric meaning of the verb *распылять* ‘to spray’ requires nouns that belong to the classes of liquids and powder-like substances (e.g. *вода* ‘water’, *духи* ‘perfume’, *химикаты* ‘chemicals’, etc.); meanwhile, the metaphoric meaning of this verb (‘to scatter, to disperse smth thus decreasing its efficiency’) combines with arguments which denote valuable resources (e.g. *деньги* ‘money’, *средства* ‘funds’, *усилия* ‘efforts’, *энергию* ‘energy’, *войска* ‘troops’, *резервы* ‘reserves’, etc.).

Halliday and Hasan (1976) point out that cohesion of a text (both within a sentence and across sentences boundaries) is provided by chains of lexical items which are in some way associated with each other in the language system, e.g.:

– *candle … flame … flicker*;
– *hair … comb … curl … wave*;
– *poetry … literature … reader … writer … style*;
– *sky … sunshine … cloud … rain* (Halliday & Hasan, 1976, pp. 285–286).

The association between the words forming such chains, according to the authors, depends on their tendency to share the same lexical environment, to occur in collocation with one another – i.e. on the relative probability with which one word tends to co-occur with another in the language system. Importantly, lexical chains are independent of the grammatical structure of the text. It will be only reasonable to suggest that metaphoric and non-metaphoric usages of verbs should participate in lexical chains which are formed by words belonging to distinctly different associative fields. Consider the associative chains in the examples from our experimental corpus:

[36] (NONMET) *Чарли пел, танцевал, < жонглировал > , делал акробатические номера*. ‘Charlie sang, danced, < juggled >, and performed acrobatic stunts.’

[37] (MET) *Очень приятно смотреть на талантливых детей, ведь они своим задором, весельем, своей непосредственностью излучают необыкновенное тепло, передают энергию, < зажигают > окружающих своим мастерством*. ‘It is a great pleasure to watch talented children: with their fervor, cheer, and ingeniousness, they radiate incredible warmth; they transmit their energy and < ignite, lit. inspire > the viewers with their mastery.’

[38] (MET) *Взбираетесь ли Вы наперегонки по стене для занятий альпинизмом или < зажигаете > на танцевальной площадке, Ваш семейный отдых никогда не будет однообразным.* ‘Whether you will choose to race up and down the climbing wall, or to < ignite, lit. to party > on the dance floor, your family vacation will never be boring.’

[39] (MET) *Брайан Трейси буквально " < взорвет > " Ваше сознание и даст мощнейшую мотивацию для ежедневных действий и достижения успеха в Вашем бизнесе.* ‘Brian Tracy will literally “ < exlplode > “ your conscience and will give you powerful motivation for daily action in order to achieve success in your business.’

1.2. Motivation behind the use of lexical co-occurrence feature

The motivation for using lexical co-occurrence as a feature for training the metaphor identification classifier bears on the extensions of the theories presented in Section 1.1 which emerge when the conceptual and semantic consideration are projected onto the surface level of their lexical manifestations in the text.

The semantic restrictions imposed on the arguments of predicates vary in the degree with which the set the boundaries of the lexical groups capable of manifesting these semantic classes.

For example, the non-metaphoric meaning of the verb *доить* ‘to milk’ will combine with the semantic class of (primarily) domesticated female animals that are capable of producing milk. The scope of nouns that cover this semantic field is rather narrow, it is almost a closed set; normally, it would include cows and goats, less frequently – sheep and mares; contexts related to certain cultures may also feature female buffalos, camels, reindeers, donkeys, and yaks.

In the case of the central literal meaning of the verb *разбавлять* ‘to dilute smth’, which is a three-place predicate, its two inanimate arguments (the Theme and the Means) belong to the semantic classes of liquids (of varying viscosity) and powder-like or granular substances. The lexical expression of these two classes ranges widely, they are potentially open classes yet they are not boundless; e.g., consider the examples collected from the first two pages of the corpus (ruTenTen11) output: *молоко* ‘milk’, *коньяк* ‘cognac’, *кофе* ‘coffee’, *сок* ‘juice’, *спирт* ‘spirit’, *вода* ‘water’, *соус* ‘sauce’, *бульон / отвар* ‘broth’, *напиток* ‘beverage’, *вино* ‘wine’, *пиво* ‘beer’, *уксус* ‘vinegar, *чай* ‘tea’, *морс* ‘fruit drink’, *сироп* ‘syrup’, *сметана* ‘sour cream’, *клей* ‘glue’, *препарат / продукт / вещество* ‘substance’, *смесь* ‘mixture’, *молотый мел* ‘ground chalk’, *перец* ‘pepper’, *мука* ‘flour’, *эмаль* ‘enamel’, *материал* ‘material’, *грунтовка* ‘primer’, *растворитель* ‘solvent’, *краска* ‘paint’, *цемент* ‘cement’, *раствор* ‘solution’, etc.

A still broader range of lexical expression is found in the primary central meaning of the verb взвесить ‘to weigh smth’. The only semantic restriction which is placed on its indirect object (the Theme) is that it must be a physical object (probably, with a small reservation that this physical object has to be commensurate with the scale of human measuring devices) – that is, one can weigh virtually everything: *пюре* ‘purée’, *пища / продукты* ‘food’, *ингредиенты* ‘ingredients’, *бананы* ‘bananas’, *колбаса* ‘sausage’, *картошка* ‘potatoes’, *мясо* ‘meat’, *сахар* ‘sugar’, *печенье* ‘biscuits’, *урожай* ‘harvest’, *кубок* ‘goblet’, *ложка* ‘spoon’, *нож* ‘knife’, *ребенок* ‘child’, *малыш* ‘baby’, *беременная* ‘pregnant woman’, *пациент* ‘patient’, *испытуемый* ‘subject (in an experiment)’, *слон* ‘elephant’, *щенок* ‘puppy’, тигр ‘tiger’, *окунь* ‘perch’, пчелы ‘bees’, рюкзак ‘backpack’, *вещи* ‘clothes’, *ботинок* ‘boot’, *ручная кладь* ‘carry-on luggage’, *ноша* ‘lit. burden’, *коробка* ‘box’, *тара* ‘packaging’, *груз* ‘cargo’, *цистерна* ‘tank, *деньги* ‘money’, *тетрадь* ‘copybook’, *товар* ‘goods’, *гиря* ‘kettlebell’, *камень* ‘stone’, *палка* ‘stick’, *дерево* ‘tree’, *пенобетон* ‘foam concrete’, *автомобиль* ‘car’, *песчинка* 'grain of sand', etc.

Beside semantic combinability pointed out by the Moscow Semantic School, some verbs possess lexical combinability (recall from Section 1.1) – when a valence of a predicate is filled with lexemes that do not share any common semantic component and form a closed set. For instance, one of the metaphoric meanings of the verb *взорвать* ‘to discontinue, lit. to explode smth’ can be realized with only one word, *тишина* ‘silence’:

[40] *Вдруг тишину < взорвал > отчаянный вопль …* ‘The silence was suddenly < interrupted, lit. exploded > by an anguished shriek.’

1.3. Motivation behind the use of morphosyntactic co-occurrence feature

1.3.1. The integrative model of lexical semantics: the Moscow Semantic School

The integrative model of lexical semantics developed by the Moscow Semantic School views word meaning as holistic unity of the semantic characteristics of a word and its syntactic arguments, along with the morphological expression of the latter arguments (Apresyan, 1995, 2009).

The number of a word's valences and the semantics of these valences are defined by the extralinguistic situation denoted by the predicate. For example, the verb *расти* 'to grow' is a one-place predicate: the only valence it features belongs to the subject (*what* grows). To compare, the verb *пилить* 'to saw' is a four-place predicate with the following valences: who, what, with what (the instrument), for what (the result). A change in the number of valences results in the change of the word's semantics; for example, the verb *арендовать* 'to rent' has five actants: the subject (*who* rents), the object (*what* is rented), the counteragent (*from whom* the object is rented), the second object (the rental), and the time period (the duration of the rent). If the last actant (the duration of the rent) is removed from the set of arguments, the meaning of the word transforms into the situation of purchasing and selling. If the first object is excluded, the resulting meaning will describe the situation of loaning; excluding the second object and the duration will shape the remaining arguments into the situation donating (Apresyan, 1995, p. 120).

The morphological expression of the arguments of a predicate can be either idiomatic or non-idiomatic. When the morphological expression of an argument is solely conditioned by the valence, the morphological form of the argument is non-idiomatic and more or less uniform across the predicates. For example the valence of cause ('because of smth / as a consequence of smth') in the majority of Russian words is expressed by prepositional and conjunctional constructions: *из-за чего-л.* iz-za$_{_prp}$ S$_{_Gen}$, *из-за того, что ... / потому что ... / по той причине, что ...* iz-za togo, chto$_{_conj}$ Clause / potomu chto$_{_conj}$ Clause / po toj prichine, chto$_{_conj}$ Clause. Similarly, the valence of goal in most Russian words is expressed non-idiomatically by means of constructions *для чего-л., с целью чего-л.* dlya$_{_prp}$ S$_{_Gen}$ / s cel`yu$_{\_conj}$ V$_{\_inf}$ / s cel`yu$_{_conj}$ S$_{_Gen}$.

Morphological expression is idiomatic when it is conditioned both by the semantics of the valence and the predicate; for instance, the valence of the object in different verbs requires different morphological forms for its arguments, cf.:

- 'to touch smth' – *дотрагиваться до чего-л.* V do$_{_prp}$ S$_{_Gen}$ and *задевать (за) что-л.* V (za$_{_prp}$) S$_{_Acc}$;
- 'to affect smth' – *влиять на что-л.* V na$_{_prp}$ S$_{_Acc}$ and *сказываться / отражаться на чем-л.* V na$_{_prp}$ S$_{_Prp}$;
- 'to occupy oneself with smth' – *заниматься чем-л.* V S$_{_Instr}$ and *работать над чем-л.*V nad_prp S$_{_Instr}$ (Apresyan, 1995, p. 121).

Thus, the same valence can receive a large variety of morphological expressions with different predicates and vice versa: the same morphological form can express various valences. For example, the valence of Instrument in Russian can be expressed by the following forms:

- *Из $S_{\_Gen}$* iz$_{_prp}$ $S_{\_Gen}$ (*стрелять из ружья* ‘to shoot a gun’);
- S*_Gen* (*курить трубку* ‘to smoke a pipe’);
- *Через S_Gen* cherez$_{_prp}$ $S_{\_Gen}$ (*просеять через решето* ‘to sift through a sive’);
- $S_{\_Instr}$ (*резать ножом* ‘to cut with a knife’);
- *С $S_{\_Instr}$* $S_{\_prp}$ $S_{\_Instr}$ (*прыгать с парашютом* ‘to parachute’);
- *На $S_{\_Prp}$* na$_{_prp}$ $S_{\_Prp}$ (*тереть на терке* ‘to grate with a grater);

At the same time, the form of the Instrumental case of nouns ($S_{\_Instr}$) can express the following valences:

- Subject (*строиться рабочими* ‘to be built by workers’);
- Object (*сорить деньгами* ‘to squander money’);
- Instrument (*резать ножом* ‘to cut with a knife’);
- Means (*мазать мазью* ‘to rub with ointment’);
- Cause (*подавиться костью* ‘to choke with a bone’), etc. (Apresyan, 1995, p. 145).

Differences in the morphological expression of arguments of one predicate can distinguish different semantic shades of the valence. To illustrate, the valence of the Instrument in the verb *шить* ‘to sew’ can be expressed by two morphological forms:

- $S_{\_Instr}$ (*шить иглой* ‘to sew with a needle’);
- На $S_{\_Prp}$ (*шить на швейной машинке* ‘to sew with a sewing machine’).

The difference lies in the relation between the Subject and the Instrument: in the first case, the position of the Instrument is not fixed, and Subject manipulates the Instrument in an unconstrained fashion. In the second case, the Instrument is stationary, and the Subject can manipulate the Instrument only in a predefined way.

To give another example, consider the verb *предлагать* ‘to suggest’ (in the meaning ‘to offer and idea or plan for someone to consider’); the valence of Content (*what* is suggested) in this verb can be expressed in the three ways:

- $S_{\_Acc}$ (*Врач предложил больному перемену обстановки* ‘The doctor suggested that the patient should move to a different environment’);
- $V_{\_Inf}$ (*предложить кому-л. поработать / пройтись* ‘to suggest doing some work / going for a stroll);
- *Чтобы + придаточное предложение* chtoby$_{_conj}$ + Clause (*предлагать, чтобы дети отдохнули* ‘to suggest that children should have some rest’).

In the first and the third case, the subject of the suggested action is the Addressee; in the second case, however, the suggested action can be addressed to both the Addressee and the Subject of suggesting: *Я предложил ему пройтись* ‘I suggested going for a stroll’ can mean either ‘I suggested that *he* should go for a stroll’ or ‘I suggested that *we* should go for a stroll’ (Apresyan, 1995, pp. 143–145).

1.3.2. Construction Grammar

The basic tent of Construction Grammar (CxG) as developed in (Brugman, 1988; Fillmore, 1985, 1988; Fillmore, Kay, & O’Connor, 1988; Goldberg, 1995; Lakoff, 2008; Lambrecht, 1994) is that constructions are the basic unit of language. Constructions are form-meaning correspondences that exist independently of particular verbs. That is, constructions themselves carry meaning, independently of the words in the sentence. Systematic differences in meaning between the same verb in different constructions are attributed directly to the particular constructions. Several constructions can be associated with a family of distinct but related senses, much like lexical polysemy. Moreover, constructions themselves are interrelated: they form complex networks.

In Construction Grammar, no strict division is assumed between the lexicon and syntax; both lexical and syntactic constructions are essentially the same type of declaratively represented data structure: both pair form with meaning.

A distinct construction is defined to exist if something about their form or meaning is not strictly predictable from the properties of their component parts or from other constructions. That is, a construction is posited in the grammar if it can be shown that its meaning and/or its form is not compositionally derived from other constructions existing in the language (Goldberg, 1995, p. 4,7).

The semantics of argument structure constructions is distinguished from the semantics of the verbs that instantiate them. Construction Grammar argues with the decompositional view on the verb’s meaning (such as ‘X CAUSES Y TO RECEIVE Z’, ‘X ACTS’, or ‘X CAUSES Y TO MOVE Z’); according to CxG, such structures correspond to *constructional meanings*. Only in the limiting case do verbs have such skeletal meanings (e.g. *give*, *do make*). As for the meaning of verbs, it involve rich frame-semantic knowledge which requires multiple aspects of complex knowledge of the world and culture. For example, consider the following (oversimplified) definitions:

– **riot:** for three or more people, acting as a group, to engage in activities outside of cultural norms in an unruly and aggressive manner, often with the intention of effecting political consequences.

– **renege:** to change one's mind after previously having made a promise or a commitment to do something (Goldberg, 1995, pp. 27–28).

Rich frame-semantic knowledge is necessary, among the other things, in order to account for novel uses of verbs in particular constructions. Consider, for example, the following expression:

– *Sam sneezed the napkin off the table*.

In order to interpret (or generate) this expression, one needs to know that sneezing involves the forceful expulsion of air. This would not be captured by skeletal decompositional entry for sneeze such as, for example 'X ACTS' (Goldberg, 1995, p. 29).

The mapping between syntax (decompositional constructions) and semantics (rich frame-semantic knowledge) is done via constructions, so that constructions and verbs interact in nontrivial ways.

The repertoire of constructions is not an unstructured set. Constructions form a network and are linked by inheritance relations which motivate many of the properties of particular constructions. The networks of constructions follow the general psychological principles of language organization:

I. *The Principle of Maximized Motivation:* If Construction A is related to Construction B syntactically, then the system of Construction A is motivated to the degree that it is related to Construction B semantically. Such motivation is maximized.

II. *The Principle of No Synonymy:* If two constructions are syntactically distinct, they must be semantically or pragmatically distinct. Pragmatic aspects of constructions involve particulars of information structure, including topic and focus, and additionally stylistic aspect of the construction such as register.

    Corollary A: If two constructions are syntactically distinct and semantically synonymous, they must not be pragmatically synonymous;

    Corollary B: If two constructions are syntactically distinct and pragmatically synonymous, they must not be semantically synonymous.

III. *The Principle of Maximized Expressive Power:* The inventory of constructions is maximized for communicative purposes.

IV. *The Principle of Maximized Economy:* The number of distinct constructions is maximized as much as possible, given Principle III (Goldberg, 1995, pp. 67–68).

Constructions that are related both semantically and syntactically are connected by asymmetric inheritance links. That is, Construction A *motivates* construction B iff B *inherits* from A. A link that recurs often throughout the grammar can be predicted to be productively applied to new cases which share the relevant factors with the existing cases. In a sense, a highly recurrent motivation link is quite analogous to a rule: the existence of one construction will predict the existence of an extension related by the productive link (Goldberg, 1995, p. 77).

When two constructions are found to be related by a metaphorical mapping, a *metaphorical extension link* is posited between them. The way the dominating construction's semantic is mapped to the dominated construction's semantic is specified by the metaphor. For example, it is argued that the resultative construction crucially involves a metaphorical interpretation of the result phase as metaphorical type of goal, cf.:

[41] Pat hammered the metal flat.

[42] Pat threw the metal off the table.

The resultative construction in Sentence [41] can be seen to be a metaphorical extension of the caused-motion-construction exemplified in Sentence [42], which involves literal caused motion. For more examples, consider:

[43] The jello *went from* liquid *to* solid in a matter of minutes.

[44] He *couldn't manage to pull himself out of* his miserable state.

[45] No one could help her as she *slid into* madness (Goldberg, 1995, pp. 81–84).

1.4. Motivation behind the use of concreteness feature

Implementation of concreteness feature in our experimental classifier builds on the groundwork of the theory of embodied and grounded cognition and the notion of primary metaphor (Barsalou, 2008, 2010; Lakoff & Johnson, 1999); for a detailed discussion of the theory see (Calvo & Gomila, 2008; Varela, Thompson, & Rosch, 2017); differences between embodied, grounded, and situated cognigion are discussed in (Pezzulo et al., 2013).

Grounded theories increasingly challenge traditional views of cognition by proposing that the conceptual representations underlying knowledge are grounded in sensory and motor systems, rather than being represented and processed abstractly in amodal conceptual data structures.

The grounded perspective offers a unifying view of cognition. It stresses dynamic brain-body-environment interactions and perception-action links as the common bases of simple behaviors as well as complex cognitive and social skills, without ontological (or representational) separations between these domains (Pezzulo et al., 2013).

In an embodied mind, the same neural system engaged in perception (or in bodily movement) plays a central role in conception. An embodied concept is a neural structure that is actually part of, or makes use of the sensorimotor system of our brains. Much of conceptual inference is, therefore, sensorimotor inference. It is the properties of the human body that contribute to the peculiarities of our conceptual system. We have eyes and ears, arms and legs that work in certain very definite ways and not in others. We have a visual system, with topographic maps and orientation-sensitive cells, that provides structure for our ability to conceptualize spatial relations. Our abilities to move in the ways we do and to track the motion of other things give motion a major role in our conceptual system. The fact that we have muscles and use them to apply force in certain ways leads to the structure of our system of causal concepts. Thus, basic level concepts depend on motor movement, gestalt perception, and mental imagery, which is carried out in the visual system of the brain. Our color concepts are intimately shaped not merely by perception as a faculty of mind but by such physical parts of our bodies as color cones and neural circuitry. Spatial-relations concepts (e.g. front and back) are not characterized by some abstract, disembodied mental capacity but rather in terms of bodily orientation (Lakoff & Johnson, 1999).

The growing body of empirical evidence increasingly suggests that simulations, situations, and bodily states play central roles in cognition. Experiments confirm that modal stimuli and simulation are utilized in perception (of shape, color, motion, space, and speech), in action coordination and conceptual processing, in memory and language comprehension, in concrete and abstract reasoning, and in social cognition (Barsalou, 2008).

Reason is not purely literal, but largely metaphorical and imaginative: is based on various kinds of prototypes, framings, and metaphors. Conceptual metaphor allows conventional mental imagery from sensorimotor domains to be used for domains of subjective experience.

Metaphors are organized as hierarchical systemic structures in which primary metaphors which are immediately entrenched in our bodily experiences, combine to produce complex metaphors just like atoms can be put together to form molecules; complex metaphors, in their turn, can be used as the basis for even more complex metaphors which occasionally give rise to novel metaphors.

Primary metaphors, from a neural perspective, are neural connections learned by coactivation. They extend across parts of the brain between areas dedicated to sensorimotor experience and areas dedicated to subjective experience. From a conceptual point of view, primary metaphors are cross-domain mappings, from a source domain (the sensorimotor domain) to a target domain (the domain of subjective experience), preserving inference and sometimes preserving lexical representation. Indeed, the preservation of inference is the most salient property of conceptual metaphors (Lakoff & Johnson, 1999).

The theory of primary metaphor can be attributed to the following four lines of thinking.

Johnson's theory of conflation in the course of learning (Johnson, 1996, 1999) argues that for young children, subjective (nonsensorimotor) experiences and judgments, on the one hand, and sensorimotor experiences, on the other, are so regularly conflated – undifferentiated in experience

– that for a time children do not distinguish between the two when they occur together. For example, for an infant, the subjective experience of affection is typically correlated with the sensory experience of warmth, the warmth of being held. During the period of conflation, associations are automatically built up between the two domains. Later, during a period of differentiation, children are then able to separate out the domains, but the cross-domain associations persist.

According to Grady (Grady, 1997), primary metaphors arise naturally, automatically, and unconsciously through everyday experience by means of conflation, during which cross-domain associations are formed. Complex metaphors are formed by conceptual blending. Universal early experiences lead to universal conflations, which then develop into universal (or widespread) conventional conceptual metaphors.

In Narayanan's neural theory of metaphor (Narayanan, 1997), "associations" made during the period of conflation are realized neurally in simultaneous activations that result in permanent neural connections being made across the neural networks that define conceptual domains. These connections form the anatomical basis of source-to-target activations that constitute metaphorical entailments.

The theory of conceptual blending by Fauconnier and Turner (Fauconnier & Turner, 2002) claims that distinct conceptual domains can be coactivated, and under certain conditions connections across the domains can be formed, leading to new inferences. Conceptual blends are the mechanism by which two or more primary metaphors can be brought together to form larger complex metaphors.

In summary, primary metaphors are part of the cognitive unconscious: we acquire them automatically simply by functioning in the most ordinary of ways in the everyday world from our earliest years and the rest of our lives we naturally think using hundreds of primary metaphors (Lakoff & Johnson, 1999).

1.5. Motivation behind the use of lexical signals of metaphor (“flag words”) and quotation marks

Andrew Goatly conducted a large-scale corpus-based study of metaphor reported in (Goatly, 1997) where he, among the other aspects of metaphoric usage, pointed out the lexical items that may function as signals of metaphoric language in text. The paragraphs below contain a short summary of Goatly’s typology of metaphor signals (or flag words):

Literally **and other intensifiers** (*actually, really, in fact, indeed, simply, fairly, just, absolutely, completely, fully, quite, thoroughly, utterly, regular*):

[46] By directing electronic waves to her brain, it may, quite ***literally***, tickle her fancy.

Hedgers and downtoners (*a touch, a bit, a little, something, in alone way, somewhat, rather, pretty much/nearly, more or less, not exactly/ precisely/ quite, little more than, almost, nearly, as near as makes no matter, virtually, practically, no more than, or whatever…or something*):

[47] There is ***something*** venomous about the hardness of this rock.

[48] It stood ***rather*** like an old farm dog.

[49] Those eyes in Ma Garrett's face were ***no more than*** reflectors.

Mimetic markers:

[50] But just now there was so much light that the very stones seemed semi-precious, ***a version*** of the infernal city.

[51] He places the capsule on her tongue and she closes her mouth. It is all a kind of ***parody*** of a communion.

Symbolism terms (*token, sign, symbol, instance, example, acme, epitome, prototype, the type, symbolical / -ly*):

[52] the stone, that ***token*** of preposterous time

[53] To put on someone else's clothes is ***symbolically*** to take on their personality.

Superordinate terms *((a) kind of/sort of, type of, form of*):

[54] I have no relish for the country; it's a ***kind of*** healthy grave.

[55] Too many rules; it was a pre-packed ***kind of*** life.

[56] The lungs functioned as a ***sort of*** carburettor.

Similes and comparisons:

[57] I've heard that cancer is ***like*** opening a bag of feathers in the wind.

[58] He chased the little boys about and made noises ***like*** a dog tormenting cows.

[59] He was as fitted to survive in the modern world ***as*** a tape worm in an intestine.

[60] He made his ***bear's*** way down the ladder, paw after paw.

[61] Their legs and arms were ***stick*-**thin.

Similes and comparisons with perception terms (*seem, appear, feel like, taste like, sound like, look like*):

[62] those parkas that ***look like*** balloons but weigh about an ounce

[63] He had a moment of fantasy when the stone ***seemed*** as soft as a pillow.

So to speak/in a manner of speaking:

[64] He was a skinny witty old negro. He was, ***so to speak***, in the semifinals of life.

But that's only the tip of the iceberg ***so to speak***. (Goatly, 1997, pp. 169–196)

Goatly (1997) and Veal (2018) discuss the role of such orthographic device as **quotation marks** in evoking non-default figurative interpretations of utterances. For example, in Sentence [65] inverted commas mark off metaphor from literal language, while in Sentence [66] they are used to create an ironic effect.

[65] some Western observers suspect that many figures had been “padded” for so long (Goatly, 1997);

[66] “cultured” gentlemen pursue ladies the way feral predators pursue their prey (Veale, 2018).

Quotation marks may perform multiple functions in text; their most common functions include marking reported speech, citations, and rare or unusual words. Quotes are also used to introduce terminological conventions and to make back-references to them; in the Russian tradition of punctuation, quotes mark named entities such as titles of books, works of art, organizations, locations, brand names, etc., e.g.:

[67] (NONMET) *В столице < напали > на музыканта из « Виртуозов Москвы ».* A musician from the “Moscow Virtuosi” (orchestra) has been < attacked > in Moscow.

[68] (NONMET) *Именно йеменская ячейка " Аль-Каиды " взяла на себя ответственность за неудачную попытку взорвать на прошлой неделе американский пассажирский самолет .* It was the Yemeni cell of “Al-Qaeda” who claimed to have staged the failed bombing attack on the US passenger jet last week.

[69] (MET) *При прохождении города нас < бомбардировали > песней « Какой изумительный мир » Луиса Армстронга в стиле техно .* As we were marching down the streets, we were < bombarded > by a techno cover version of Louis Armstrong’s “What a Wonderful World” .

However, preliminary examination of the corpus leaves a strong impression that metaphoric discourse encourages speakers to use quotes to mark non-figurative language (as in Sentences [65]-[66]). Consider the metaphoric examples:

[70] *По словам Обамы , сейчас медиа - индустрия круглосуточно и семь дней в неделю " < бомбардирует > " человека данными самого различного рода , причем , многие из них не являются достоверными .* According to Obama, today’s media industry on the 27/7 basis “ < bombards > ” people with all kinds of data, and much of these data are not trustworthy.

[71] *Получается, что комбинат, вгоняя в долги муниципальное предприятие, " < доит > " своих же рабочих.* It turns out that the factory is driving the municipal company into debt while “ < milking > “ (lit. profiteering on) its own employees.

[72] *Вот какими категориями " < жонглирует > " современная реклама!* It is this the kind of concepts that contemporary advertising " < juggles > " (lit. manipulates) with!

[73] *Я постаралась " < причесать > " перевод до такой степени, что при вдумчивом чтении все становится понятно.* I have tried to " < comb > " (lit. improve) the translation so that everything became clear to a thoughtful reader.

[74] *А белый цвет в интерьере " < разбавляют > " только очень светлое дерево и изменяющиеся оттенки неба .* As for the white color of the interior design, it is " < dissolved > " (lit. alternates with) only by very light-toned woodwork and the opalescent shades of the sky.

2. Data preprocessing and the context windows

The tokenization, lemmatization, as well as the morphological and syntactic parsing of the corpus were produced with Ru-syntax, a pipeline combining a morphological and a syntactic analyzers (Droganova & Medyankin, 2016). Tokenization, lemmatization and the first stage of morphological analysis in Ru-syntax are provided by Mystem (Segalovich, 2003); on the second stage the morphological information is disambiguated using TreeTagger (Schmid, 1994) with the parameter model trained on the disambiguated part of Russian National Corpus (Fenogenova, Kayutenko, & Dereza, 2015). The accuracy of the Ru-syntax morphological analyzer measured in a strict sense (i.e., one missing or misplaced tag for a token is a miss, full match tag-by-tag is a hit) is 85.5%. Syntactic annotation Ru-syntax is provided by MaltParser (Nivre et al., 2007) working in parse mode. The parsing model was trained on SynTagRus , a syntactically annotated corpus of Russian (Boguslavsky, 2014); the quality of the parsing model is 83.7% and 89.6% according to labeled and unlabeled attachment scores, respectively. We used the syntactic relations obtained with Ru-syntax in the initial experiments with windows of various sizes (including the window composed of the target verb's syntactic arguments); however, the latter type of window proved inefficient, and the syntactic annotation of the corpus was not used in the subsequent experiments.

The window-dependent features described in the next sections (i.e. the semantic, the lexical, and the morphosyntactic features) were computed:

a) on the fixed context windows of the sizes 2, 3, 4, and 5;

b) on the unfixed-size window equivalent to the length of the full sentence;

c) on the set of the syntactic arguments of the target verb (its direct dependencies and some of their secondary projections).

As we will show below (Section 3.1), the best performance of the classifier was obtained on the windows of the maximum sizes (window 5 and full sentence).

Only content non-stopwords were included into the semantic and the lexical windows; as for the morphosyntactic windows, they were comprised of all the lemmas found within a given window, including prepositions and punctuation marks.

3. The feature set

3.1. Distributional semantic features

3.1.1. The word-embedding models

Our distributional semantic features are based on word-embedding models. The hypothesis underlying distributional models of word meanings is the so-called distributional hypothesis: the idea that "Words that occur in similar contexts tend to have similar meanings" (Turney & Pantel, 2010). Harris (1954), in the tradition of the structural linguists, proposed that linguistic units, such as parts of speech, could be identified from corpora by observing the contexts in which the units occur. Perhaps the historical work most closely related to modern distributional semantics is that of Firth (1957), who was interested in the notion of collocation and how the distributional contexts of a word could be used to explain its behaviour. Firth was also interested in different word senses, arguing that the different senses of an ambiguous word could be revealed by looking at the different contexts in which the word occurs (Clark, 2015).

The distributional representation of a lexical item is a distributional vector representing its co-occurrences with linguistic contexts in a vector space, typically of several hundred dimensions. A distributional semantic model is a particular configuration of the following parameters: the selection of target lexemes, the definition of context type, the choice of weighting scheme, the application of dimensionality reduction, and the choice of a vector similarity metric (Lenci, 2018).

We experimented with two pre-trained models by A. Kutuzov and E. Kuzmenko, which are freely available for download from the RusVectōrēs website (Kutuzov & Kuzmenko, 2016; 'RusVectōrēs', n.d.). Both models were trained with the word2vec Continuous Skip Gram algorithm using the Gensim library (Rehurek & Sojka, 2010). Word2vec is a shallow, two-layer neural network; the Continuous Skip Gram algorithm tries to maximize classification of a word based on another word in the same sentence. More precisely, it uses each current word as an input to a log-linear classifier with continuous projection layer, and predict words within a certain range before and after the current word (Mikolov, Chen, Corrado, & Dean, 2013).

The first of the word-embedding models we use – **the WikiRNC model** – was trained with vector dimensionality 300 and window size 2 on the joint corpus of Russian Wikipedia and the Russian National Corpus ('Russian National Corpus', n.d.), with the total of 600m tokens;

The second model was trained on a much larger *WaC* (Web as Corpus) collection, the Araneum Russicum Maximum corpus (Benko & Zakharov, 2016), of about 10bn tokens, with vector dimensionality 600 and the window size of 2. In the subsequent sections it will be referred to as **the Araneum model**.

3.1.2. The semantic similarity measure

In order to capture both the semantic incongruity and cohesion evoked by metaphoric occurrences (as outlined in Section 1.1) we applied the measure of semantic similarity obtained by means of word-embeddings. We proceed from the intuition that a metaphoric verb will be semantically deviant from its context window, affecting the mean semantic similarity between the words in the window in a negative way, whereas a literally used verb will belong to the same conceptual domain as its context words, making the contextual sub-space denser and adding to the mean similarity .
The measure of semantic similarity was initially proposed by Newman et al. (2010) as a means to evaluate the coherence of topics generated by topic modelling algorithms; semantic similarity measure also proved efficient in detecting lexical errors in learners' English (Herbelot & Kochmar, 2016). Thus, application of this measure renders the task of metaphor identification as a subcase of lexical anomaly detection in linear context.
The semantic similarity of tokens within the context window is calculated as the following:

$$Sim_{Win} = Mean\{Sim(w_i, w_j) \mid w_i, w_j \in Win\} \quad (1)$$

$$SimV_{Win} = Mean\{Sim(w_i; w_j) | w_i, w_j \in Win, w_i \neq verb, w_j \neq verb\} \quad (2)$$

$$SimDiff_{Win} = Sim_{Win} - SimV_{Win} \quad (3)$$

where *Sim* is the semantic similarity in the distributional semantic space, and *Win* is the context window around the target verb: a linear window in the case of linear context, or the list of syntactic arguments in the case of the syntactic arguments' context.

3.1.3. The augmented semantic features

When a sentence in our corpus features a low-frequency word (or words) that is missing from the word-embeddings model, its measure of semantic similarity with its environment equals to zero. We moderated this effect by replacing the unavailable similarities by the mean of all the similarity measures in the current context window:

$$Sim_{w_i, w_j} = Mean\{Sim_{Win} \mid w_i \notin Sim_{win}\} \qquad (4)$$

Thus, we experimented with two types of distributional semantic features – the non-augmented ones, where the words which were absent from the word-embeddings models were merely disregarded, and the non-augmented features described above. We compare the performance of the two classifiers below in Section 5.4.

3.2. Lexical co-occurrence features

3.2.1. The lexical measure of metaphor association

Measures of association are statistical measures which are used in linguistics to assess the degree of attraction between words and other units. Association measures are based on information about the relative frequency of co-occurrence of words and other units, i.e. they account for the likelihood of words and other units occurring together by chance. Different measures of association are computed using all or at least some of the values from a 2-by-2 matrix of contingency shown in Table 5.

Table 4. Contingency matrix: a general view

	unit *Y*	**¬ *Y* (all other units)**
unit *X*	*a*	*b*
¬ *X* (all other units)	*c*	*d*

For two words (or any other units) *X* and *Y*, these frequencies can be defined as follows:

- *a* corresponds to the number of co-occurrences of *X* and *Y*;
- *b* is the number of occurrences of *X* without *Y* (in all other contexts), i.e. the total number of occurrences of *X* minus *a*;
- *c* corresponds to the sum frequency of all other target words that co-occur with *Y*, i.e. the total frequency of *Y* minus *a*;
- *d* is the frequency of all units except *X* that occur with all units except *Y*. In practice, this is the total number of words, sentences or other units in the corpus minus *a*, *b* and *c*. A more specific definition of this total number depends on the research question (Levshina, 2015).

In the framework of our experimental setup, the contingency matrix for computing lexical co-occurrence features contained the following frequencies (Table 6): for each lexeme *X*, we recorded the raw frequencies of its occurrences in the metaphoric and the non-metaphoric subcorpora, along with the total of the frequencies of all other lexemes in each subcorpus. The lexemes of the target verbs were excluded from the analysis.

Table 5. Contingency matrix in the metaphor classification experiment

	metaphoric subcorpus, raw frequencies	**non-metaphoric subcorpus, raw frequencies**
lexeme *X*	*a*	*b*
¬ *X* (all other lexemes)	*c*	*d*

There are dozens of possible measures of association proposed in the literature (e.g. Evert, 2005). In our classification experiment for metaphor identification we had run preliminary tests with several association measures, including:

- weirdness (Ahmad, Gillam, & Tostevin, 2000);

– the extension of Student's t-test proposed in (Manning & Schütze, 1999, pp. 167–168);
– log likelihood (Cressie & Read, 1984);
– Kullback-Leibler Divergence (Kullback & Leibler, 1951).

The best results in the test runs of the classifier were obtained with the ΔP metric (Levshina, 2015, p. 232) which was used in all the subsequent experiments to compute the co-occurrence-based features. ΔP was originally introduced in psychological research to measure the relationship between cue and response (Allan, 1980); more recently, ΔP was used in linguistic constructionist studies by Ellis (2006) and Ellis and Ferreira-Junior (Ellis & Ferreira-Junior, 2009). ΔP is a unidirectional (asymmetrical) contingency-based measure which is computed according to Equation 5:

$$\frac{a}{a+c} - \frac{b}{b+d} \tag{5}$$

In terms of our contingency matrix of co-occurrence between lexemes, on the one hand, and the metaphoric and the non-metaphoric subcorpora on the other (*Table 6*), Equation 5 translates as follows: *a* is the number of occurrences of a lexeme in the metaphoric subcorpus, *b* is the number of occurrences of the lexeme in the non-metaphoric corpus, $a+c$ is the size of the metaphoric subcorpus, and $b+d$ is the size of the non-metaphoric subcorpus.

The ΔP measure computed according to Equation 5 on the data from the contingency matrix presented in Table 6 will be referred to in the subsequent sections of this thesis as **the lexical measure of metaphor association,** and the scores produced with this measure will be referred to as **indexes of lexical metaphor association**.

3.3. Morphosyntactic co-occurrence feature

The vectors of morphosyntactic co-occurrence were computed with the same statistical measure and vectorization method as for the lexical co-occurrence feature described in Section 3.2: for each word in the corpus we computed **the indexes of morphosyntactic metaphor association** applying the ΔP metric to the morphological tags generated by Ru-syntax (see Section 2); the morphosyntactic vectors were computed as the mean of the metaphor association scores of the words in a context window (also the association scores of morphosyntactic bigrams and trigrams, see below). The morphological tags of the target verbs were included in the context windows.

Since in Russian verbs and nouns are the parts of speech which govern the morphological forms of their dependencies, we tested three different configurations of morphological characteristics of nouns and verbs, which vary in the fullness of representation of morphological information:

1. The ***only_POS*** configuration included only the indication of the part of speech – both for verbs and nouns;
2. The ***short*** configuration included the morphological characteristics of verbs and nouns which are syntagmatically independent of the morphological form of their dependents, i.e. the characteristics that are not conditioned by the agreement between the head and dependencies:
   - for nouns – part of speech, animacy, case;
   - for verbs – part of speech, aspect, tense, mood;
3. The ***full*** configuration included all of the available morphological characteristics of verbs and nouns:
   - for nouns – part of speech, gender, animacy, case, and number;
   - for verbs – part of speech, aspect, tense, number, mood, gender, and person;

Prepositions and punctuation marks in all the configurations were represented by their lemmas; all the other parts of speech were always represented by their POS tags.

In the preliminary runs of the classifier we experimented with five combinations of morphological configurations:

- verb *only_POS* + noun *only_POS*;
- verb *full* + noun *full*;
- verb *short* + noun *full*;
- verb *full* + noun *short*;
- verb *short* + noun *short*.

Besides, we compared the performance of the classifier on unigrams, bigrams, and trigrams of tags: while bigrams and trigrams are expected to capture the linear sequences of morphosyntactic tags, unigrams are supposed to render each context as a bag of morphosyntactic characteristics of the words comprising it.

The results of the comparison of various configuration and n-grams are reported in Section 5.4.

3.4. Concreteness feature

We use automatically generated concreteness scores; in computing these scores we adopt the approach implemented by Turney, Neuman, Assaf, and Cohen (2011) to obtain the concreteness ranking of approx. 114,500 English words. The authors began by automatically selected 40 concrete and 40 abstract paradigm words from the MRC Psycholinguistic Database Machine Usable Dictionary (Coltheart, 1981). After that, they calculated the abstractness of a given word by the sum of its similarity with twenty abstract paradigm words minus the sum of its similarity with twenty concrete paradigm words; to measure the similarity between words they used a latent semantic analysis (LSA) model trained on a 50-billion web-crawled corpus.

In our experiment, we used only concrete paradigm words; we drew our concrete paradigm nouns from the semantically annotated dataset of the thesaurus 'Open semantics of the Russian language' (Kulagin, 2017/2017, 2018). The ontology of the semantically annotated dataset is a hierarchical tree of semantic categories, e.g.:

- ENTITY: PHYSICAL:ORGANIC:ROLE:HUMAN;
- ENTITY:PHYSICAL:CONSTRUCTION:ROLE: TRANSPORT;
- ENTITY:ABSTRACT:ROLE: INFORMATION, etc.

Each word in the thesaurus is assigned a score indicating the probability that the word belongs to a given ontological category. All words are assigned a probability generated by a distributional semantic algorithm; besides, some of the words have a probability evaluation from one or two human annotators.

For our concrete paradigm we manually selected monosemous 515 nouns with the role tag 'THING' which were referred to this class both by the automatic scorer and two human annotators with probability <0.9. The full list of our concrete paradigm words is available in Appendix 2.

In order to have an interpretable account of the concrete paradigm words we clustered their vectors obtained with the Araneum word-embedding model (see Section 3.1.1) using scikit-learn's KMeans algorithm (the value of *k* was adjusted manually). The clustering with *k*=25 split the paradigm list into 24 roughly defined thematic groups, some of which are:

- **clothes:** *свитер, сарафан, сорочка, рубашка, футболка, чулок, шаль, шарф, шуба, кимоно, пижама, мини-юбка, куртка, пальто, пиджак, кофта, корсет, платье, пуховик, комбинезон, полупальто, колготки, майка, пуловер, жилет, юбка, ветровка, безрукавка*, etc.
- **tools:** *топор, электропила, электродрель, стеклорез, степлер, фуганок, тяпка, шуруп, циркуль, стамеска, шило, железяка, скребок, бритва, нож, плуг, железка,*

перфоратор, паяльник, гвоздь, кусачки, гвоздодер, лобзик, лопата, газонокосилка, отвертка, вилы, монтировка, мотыга, ножовка, etc.

- **miscellaneous, predominantly household, articles:** *наперсток, крюк, гарпун, гребенка, медаль, щеколда, спичка, подкова, плетка, коромысло, шпингалет, крючок, жетон, гвоздик, прищепка, замочек, веретено, шпулька, защелка, булавка, бирка, фитиль, рукоятка, пряжка, пуговица, болт, кандалы, стремя, кегля, уключина,* etc.
- **musical instruments:** *бас-гитара, ксилофон, электрогитара, арфа, гитара, гармошка, валторна, гобой, аккордеон, банджо, барабан, флейта, контрабас, фортепиано, рояль, виолончель, баян, тромбон, фортепьяно, мандолина, саксофон, кларнет, скрипка*, etc.
- **weapons and munitions:** *миномет, граната, ружье, двустволка, винтовка, рогатка, гранатомет, кобура, авиабомба, револьвер, нагайка, наган, сабля, пистолет, шпага, кортик, огнемет, пулемет, дробовик, дуло,* etc.

On the basis of the concrete paradigm list we computed **concreteness scores** for each notional word of the corpus (i.e. for nouns, verbs, and adjectives): we used the Araneum word-embedding model to find the ten semantically nearest words and took their mean:

$$\boldsymbol{Concr} = \boldsymbol{Mean}(\boldsymbol{Neighbor1} + \cdots + \boldsymbol{Neighbor10}) \tag{6}$$

With this approach to generating concreteness scores, some of the top concrete words in the resulting ranking are: *торшер, жаровня, мохеровый, расческа, флешка, каска, жабо, аккомпанемент, веревка, зажигалка, нагрудник, треуголка, мешочек, пианист, посуда, гитарист, винт, бурнус, карандаш, тюль, ложечка, кисточка, пеленка, шпага, проигрыватель, погон, сервиз, скатерть, копье, огнемет, салфетка, мех, зеркалка, подошва, скрипач, газонокосилка, наливать, клинок, утюг, тазик, болт, видеокамера, пушка, булавка, сундук, ножнички, оркестр, защелка, волынка.*

The least concrete words (i.e. the most abstract ones) are: *повлиять, способствовать, содействовать, установление, принятие, сформировываться, рассматриваться, формироваться, развитие, обсуждаться, влиять, являться, повышение, основываться, организованный, согласоваться, низация, вовлечение, ответственный, отношение, активизировать, взаимодействие, формирование, предпосылка, сохранение, противоречить, возможный, возрастать, пониматься, влияние, пересмотр, незамедлительный, совместный, целое, развиваться, единодушный, социально.*

The full concreteness ranking of 17,073 words is available for download at: https://github.com/yubadryzlova/metaphor_dataset_20_verbs.git

The obtained concreteness scores have not been subject to manual evaluation and verification by human assessors; such an endeavor would require an elaborate psycholinguistic concreteness evaluation protocol and a carefully designed experiment – which should be addressed in future research.

The vectors of contexts given as the input to the classifier were computed as the mean of the concreteness scores of the words within the context window, including the target verb.

3.5. Flag words and quotation marks features

Manual analysis of the experimental corpus revealed the following set of five words that could be expected to function as plausible lexical signals of metaphor (‘flag words’): *буквально* ‘literally’, *как будто*, *словно* ‘as if’, *т.е.* ‘i.e.’, and *подобно* ‘like’. Consider the examples from the experimental corpus:

[75] (MET) *В данной ситуации мы* ***буквально*** *< бомбардировали > жалобами и обращениями различные инстанции .* In that situation we **literally** < bombarded > various agencies with our complaints and appeals.

[76] (MET) *Регулярное и длительное употребление спортивных энергетиков существенно повреждает зубную эмаль , позволяя кислотам* ***буквально*** *< съедать > гладкое покрытие зуба .* Regular and lasting consumption of energy drinks for athletes seriously damages the tooth enamel and allows acids to **literally** < eat up (lit. erode) > the outer surface of the tooth.

[77] (MET) *Немецкие танки* ***буквально*** *< утюжили > окопы роты.* German tanks were **literally** < ironing (lit. crushing with their tracks) > the trenches of the infantry.

[78] (MET) *Песни совсем не < греют > ,* ***словно*** *находятся во власти холодной атмосферы .* The songs to not < make you feel warm > at all, **as if** they are under the influence of a cold atmospheric wave.

[79] (MET) *Канат ,* ***словно*** *ножовка , < пилит > плечо и шею .* The rope is < sawing (lit. chafing) > me the shoulder the neck **as if** it were a hand saw.

[80] (MET) *Лишь изредка его* ***словно*** *< подхватывало > какой-то волной - и переносило вперед на очередные несколько сантиметров.* Only at times he felt **as if** he were < seized > by some kind of wave – and carried forward by the next few centimeters.

[81] (MET) *Не знаю, что тут было, но со стороны это выглядит так,* ***как будто*** *гигантской гребенкой < причесали > сосновые локоны острова.* It is hard to tell what exactly happened here, but to an outside eye it looks **as if** a gigantic comb has < combed > the pine-tree locks of the island.

As for quotation marks, examination of the corpus allowed us to reveal the following set of characters which occur in the corpus and function as quotation marks: " … " (straight inverted commas), « … » (angular quotation marks), ' … ' (straight single quotes), "" … "" (straight double inverted commas), and * … * (asterisk). Note that this set contains no curly quotation marks (or ‘smart quotes’) – “ “ and no bi-directed pairs with upper-lower alignment, i.e. such pairs where the opening quote is located at the upper margin of the line and the closing quote is aligned with the lower margin, e.g.: “ … „. We presume that such variants may not be present in the corpus for two reasons: they may not be popular with users because the default settings of the currently most widely used word-processors are set to straight quotes; besides, curly quotes and bi-directed pairs might have been standardized and replaced with straight variants during the preparation of the ruTenTen corpus from which sentences were drawn to compile our experimental corpus.

Thus, for each contextual window, we generated the three kinds of vectors:

- One-hot encoded vectors of flag words;
- The vectors of the distance between the flag word(s) and the target verb (in tokens);
- The vectors of the distance between quotation marks and the target verb.

4. Experimental setup

The metaphor identification task was formulated as sentence-level binary classification: by looking at a sentence the classifier was to predict whether the sentence belonged to the metaphoric or the non-metaphoric class.

In total, we conducted 21 experiments: one separate experiment with the dataset of each of the 20 target verbs, and one experiment with the large combined dataset of all the 20 verbs.

We used the Support Vector Machine (SVM) classifier with linear kernel[3]. SVM is a discriminative classifier: it finds a hyperplane that divides the classes from each other and tries to maximize the margin around this hyperplane. The parameters of the classifier were tuned using grid search in the test runs on the distributional semantic features. The best results were achieved with the following set of parameters: penalty = 'l2', loss = 'squared_hinge', dual = False, C = 1000, multi_class = 'ovr'; this set of parameters was used in all the subsequent experiments.

Each experiment was run using 5-fold cross-validation.

In each experiment we implemented a total of 26 models, i.e. different one-, two-, three-, four-, and five-feature combinations from the feature set introduced in Section 3 of the present chapter.

5. Results

5.1. Evaluation of alternative parameters of the features

The alternative parameters of the distributional semantic, and morphosyntactic co-occurrence features, which were presented in Sections 3.1, 3.2, and 3.3 were compared in test runs of the corresponding classifiers where their performance was evaluated in terms of classification accuracy. The comparison was carried out in large aggregate terms – as the mean accuracy across the 20 datasets of the target verbs.

Between the two pre-trained distributional semantic models – the one trained on the combined corpus of Russian Wikipedia and the Russian National Corpus (600m tokens), and the model which was trained on the *WaC* Araneum corpus (10bn tokens) – slightly better results were obtained with the Araneum model. In the subsequent sections of this thesis we will present the performance of the classifier trained on this model when discussing the results of the distributional semantic classifier.

When comparing augmented and non-augmented distributional semantic vectors (i.e. those where the words that were absent from the pre-trained word-embedding models were filled with the mean value of all the words in the context window, and those where such substitution was not applied, respectively), no marked difference was observed in their performance. Therefore we chose to continue experiments with the non-augmented vectors on the account of their lesser complexity and computational cost.

[3] LinearSVC, as implemented in scikit-learn (Pedregosa et al., 2011).

Comparison of the five combinations of the *only_POS*, the *short* and the *full* configurations of morphological tags (see Section 3.3) revealed that demonstrably higher accuracy was achieved with the morphologically informed combination in which both nouns and verbs carried their *full* tags: verb *full* + noun *full*.

The morphosyntactic classifier which was trained on unigrams of morphological tags by far outperformed the classifiers trained on bigrams and trigrams. Analysis revealed that the bigram and trigram-trained classifiers failed due to the sparsity of the input data: the number of unique linear morphosyntactic patterns turned out to be very large, while the number of their occurrences was low.

When reporting the performance of the morphosyntactic classifier in the subsequent sections we will be presenting the results of the classifier which was trained on unigrams of full morphological tags.

5.2. Window sensitivity

All the window-dependent features (the distributional semantic, the lexical and morphosyntactic co-occurrence vectors) have proved to be sensitive to the size of the context window. Figure 1 demonstrates correlation between the classification results (accuracy) on the lexical features and the size of the window for three the verbs (*bombardirovat* 'to bombard smth', *otrubit* 'to hack (smth) off', and *prichesat* 'to comb smb/smth') which demonstrate a downward, an upward, and a flat dynamics.

Obviously, this behavior is connected with the distances at which the lemmas with conspicuous association scores occur in relation to the target verb.

For example, *otrubit* 'to hack smth off' best performs on the set of the syntactic arguments; this is due to the high frequency of the metaphoric intransitive construction which serves to introduce direct speech, e.g.:

[82] (MET) *Нет, — <отрубил> Керк. — Деньги должны быть выиграны сегодня.*

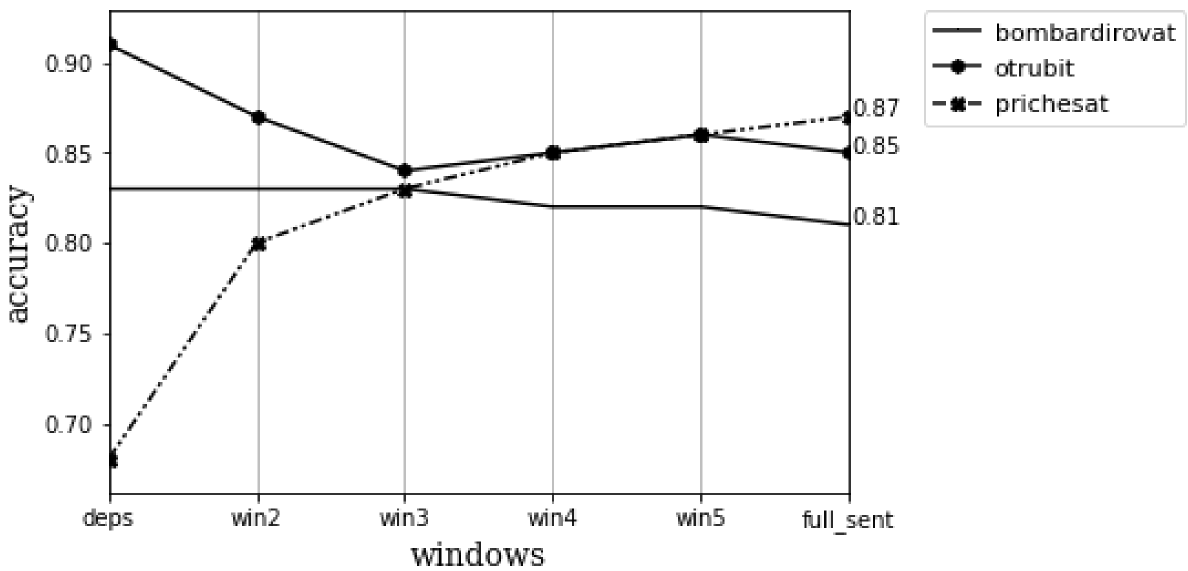
'No', Kirk <cut off> (lit. responded abruptly), 'the money must be won today'.

The verb in this construction has only one syntactic argument, the subject, which is typically a person's name. Proper names are low-frequency lemmas, and therefore they will have low metaphor association scores. Whereas the non-metaphoric meaning tends to co-occur with higher-frequency arguments on a much more regular basis (e.g. the direct objects filling the valence of the Instrument: *отрубить мечом / топором / саблей (шашкой) / штыком* 'to hack smth with a sword / ax / saber / bayonet'; also, the direct objects which fill the valence of the Patient: *отрубить голову, палец, язык, кисть, часть, кусок* 'to hack off one's head, finger, tongue, hand, part, chunk, etc.). This contrast imparts high predictive power to the model based on the syntactic dependents of *otrubit*. Using linear windows, especially larger ones, introduces excessive noise into the model and disorients the classifier.

In the case of the verb *prichesat* 'to comb smb/smth', the highest score of metaphor association belong to the lexeme *гребенка* 'comb': it occurs much more frequently in metaphoric than in non-metaphoric contexts due to the high frequency of the idiom *причесать всех под одну гребенку* 'to comb everyone with the same comb' lit.'to make everything / everyone equal, disregarding their individual features'. The word гребенка does not belong to the set of the verb's syntactic dependents, therefore which operates on them performs poorly. As the contextual window increases and spans over the idiomatic expression, the performance begins to increase.

In the aggregate terms the two large-size windows (full_sent and win5) perform on a par with each other. We chose the full-sentence window for our subsequent experiments since full sentences allow of greater interpretability.

Figure 1. Correlation between the accuracy of classification and the size of the context window (lexical co-occurrence features).



'*Deps*' – the set of the verb's syntactic arguments; '*win2*' – '*win5*' – windows of the sizes 2 – 5; '*full_sent*' – window of the full sentence length.

5.3. Inefficient features

The vectors based on lexical signals of metaphor (flag words) and quotation marks (recall from Section 3.5) proved inefficient as features for a classifier. The accuracy of the classifier trained on the flag words vectors hovered just above the chance level in all the 20 datasets of individual target verbs and the combined dataset of the 20 verbs. Post hoc analysis revealed that the reason of such feeble performance lies in the extreme sparsity of the data: occurrences of flag words are highly infrequent in the corpus, as demonstrated in Table 7.

Table 6. Raw frequencies of flag words by class (MET | NONMET)

verb / flag_word	***буквально*** **iterally**	***как будто*** **as if**	***подобно*** **like**	***словно*** **as if**	***т.е.*** **i.e.**
bombardirovat	**10 | 2**	0 | 0	0 | 1	0 | 0	1 | 0
doit	0 | 0	1 | 0	0 | 0	0 | 1	2 | 1
gret	0 | 0	0 | 0	1 | 0	2 | 0	0 | 1
napadat	2 | 0	1 | 0	0 | 0	3 | 0	0 | 0
ochertit	0 | 0	1 | 0	0 | 0	0 | 0	0 | 0
otrubit	0 | 0	0 | 0	1 | 0	1 | 0	0 | 0
pilit	0 | 0	0 | 1	0 | 0	1 | 0	0 | 2
podkhvatyvat	0 | 0	0 | 0	0 | 0	3 | 4	0 | 0
prichesat	0 | 1	1 | 2	0 | 0	0 | 1	0 | 0
raspylyat	0 | 0	0 | 0	0 | 0	0 | 0	0 | 0
razbavlyat	0 | 0	0 | 0	0 | 0	0 | 0	0 | 3
syedat	2 | 0	1 | 0	0 | 0	1 | 0	0 | 0
trubit	0 | 0	1 | 0	1 | 0	1 | 4	0 | 0
ukolot	0 | 0	0 | 1	0 | 0	**12 | 1**	0 | 0
utyuzhit	3 | 0	0 | 1	0 | 1	1 | 0	0 | 0
vykraivat	0 | 0	0 | 0	0 | 0	0 | 0	0 | 1
vzorvat	**20 | 0**	0 | 0	0 | 0	1 | 0	0 | 0
vzvesit	0 | 0	0 | 0	0 | 0	0 | 0	0 | 0
zazhigat	3 | 1	0 | 0	0 | 0	0 | 0	0 | 0
zhonglirovat	2 | 1	1 | 0	1 | 0	5 | 4	0 | 0

Nevertheless, analysis of the distribution of flag words in the corps leads to several valuable observations: note that the relatively high frequency of the flag word *буквально* in the metaphoric dataset of the verbs *bombardirovat* 'to bombard' (10 occurrences) and *vzorvat* 'to blow smth up' (20 occurrences), and the 12 occurrences of *словно* 'as if' in the metaphoric dataset of *ukolot* 'to prick'. *Ukolot* is regularly used in the simile-based innovative metaphor where *slovno* 'as if' acts as the operator of comparison which can involve two types of mapping: 1) a sudden mental state is likened to the physical action of a pricking instrument 2) the cause of a physical state is likened to the action of a pricking instrument; cf.:

[83] (MET) *Внезапно её словно < уколола > какая-то мысль.* ‘Suddenly she felt as if some thought had pricked her.’

[84] (MET) *мальчишка сразу возьми и заплачь , да так кричит , словно иглой укололи.* ‘The boy burst out weeping, screaming as if somebody pricked him with a needle.

The function of *буквально* ‘literally’ in metaphoric contexts, as Goatly (1997, p. 174) describes it, is ‘a means of strengthening the metaphorical contradiction’. Presumably, the metaphoric contexts of *bombardirovat* and *vzorvat* attract this intensifier because the central literal meanings of these verbs denote highly violent physical action as it is, and ‘literally’ is aimed to reinforce the mapping between the Source and the Target, painting a vivid picture of bombings and explosion in the recipient’s mind:

[85] *Иначе зачем производители* ***буквально*** *< бомбардируют > рынок подобными моделями ?* Otherwise, why should the makers be literally < bombarding > the market with such models?

[86] *Владимир Васильевич, ваша отставка с должности посла* ***буквально*** *< взорвала > информационное пространство.* (Mr.) Vladimir Vasilyevich, your resignation from the office of ambassador has literally < exploded > the media scene.

[87] ***Буквально*** *< < взорвали > > зал в прямом и переносном смысле герои 90-х* **–** *группа «Технология».* A 1990s idol, the “Technology” band, literally < exploded > the concert hall **–** both in the figurative and the literal meaning of this word.

Note the interplay of the markers of metaphoric and non-metaphoric meanings in Sentence [87] (‘literally’ and ‘both in the figurative and the literal meaning of this word’) which are intentionally (and possibly excessively) used to simultaneously evoke both meanings.

The accuracy of classification with the vectors based on the frequency of quotation marks in the sentence and their distance from the target verb differed across 20 datasets, from just above the chance level to moderately high 0.7-0.69 for the verbs *prichesat* ‘to comb smb / smth’ and *syedat* ‘to eat smth up’ (see Table 8).

Table 7. Accuracy of classification with the quotation marks vectors

target verb	accuracy	target verb	accuracy
bombardirovat	0,58	**razbavlyat**	0,55
doit	0,63	**syedat**	**0,69**
gret	0,55	**trubit**	0,53
napadat	0,52	**ukolot**	0,54
ochertit	0,53	**utyuzhit**	0,59
otrubit	0,51	**vykraivat**	0,55
pilit	0,65	**vzorvat**	0,62
podkhvatyvat	0,56	**vzvesit**	0,55
prichesat	**0,7**	**zazhigat**	0,6
raspylyat	0,55	**zhonglirovat**	0,55

Once again, these differences in performance correlate with the density of the data: Table 9 shows the frequencies of pairs of quotation marks in the datasets of the 20 target verbs, and the datasets of the verbs *prichesat* and *syedat* contain the highest number of these orthographic devices.

Table 8. Raw frequencies of quotation marks(paired) by class

target verb	MET	NONMET
bombardirovat	57	26
doit	93	19
gret	54	17
napadat	22	14
ochertit	15	8
otrubit	42	28
pilit	64	6
podkhvatyvat	59	15
prichesat	**138**	18
raspylyat	29	9
razbavlyat	38	15
syedat	**171**	30
trubit	63	30
ukolot	36	16
utyuzhit	69	18
vykraivat	17	5
vzorvat	81	24
vzvesit	49	15
zazhigat	54	18
zhonglirovat	68	27

We hypothesize that the high attractivity of the verbs *prichesat* and *syedat* could be related with the high agentive power of the main arguments of their central literal meanings: *prichesat* ‘to comb smb' requires an animate Patient, while *syedat* 'to eat smth up' can be used only in the construction with an animate Agent. When these two verbs are used metaphorically, it creates the effect of personification which is felt by people producing such utterances, which urges them to emphasize this effect with quotation marks more lavishly. Consider the examples from the experimental corpus:

[88] *Эта книга поможет вам « причесать » ваши гениальные мысли .* 'This book will help you to "comb" (= to brush up) your brilliant ideas.'

[89] *Почему , все кто соприкасался с этим делом , пытались « причесать » показания виновных ?* 'Why did every person involved in this case try to "comb" (lit. clean up) the testimony given by the accused?

[90] *Не « съедают » ли высокие тарифы прибавки к пенсиям ?* Don't the high prices "eat up" (lit. offset) the pension rise?

[91] *Считается , что ржавчина в среднем « съедает » до 0,1 мм неоцинкованного металла в год .* It is considered that rust on the average "eats up" (lit. corrodes) up to 0.1 mm of non-galvanized metal per year.

A similar, but a less pronounced effect of agentivity can also be observed in the verb *doit* 'to milk' in Table 9: this verb's central literal meaning also requires an animate Patient (the animal which is milked), and it has the third largest number of quotation marks in its metaphoric subcorpus (see, for instance, Sentence [71] above).

Overall, despite the marginally low utility of flag words and quotation marks and the fact that they cannot be used as standalone features to train a classifier, they seem to possess certain value. This finding can be used in future experiments. Flag words could probably be used to augment the lexical co-occurrence vectors by assigning higher weights to the occurrences of such lexemes. Quotes could be used to boost the morphosyntactic co-occurrence vectors with higher weights attached to the occurrences of these punctuation marks.

5.4. Classification results

5.4.1. Comparison of one-feature models across datasets

Table 10 demonstrates the results of the classifiers trained with one type of feature – distributional semantic, lexical co-occurrence, morphosyntactic co-occurrence, or concreteness – within each of the 20 individual verb datasets, as well as on the combined dataset of the 20 verbs; the table also presents the mean results of the 20 datasets, and the standard deviation across the 20 datasets for each model (the two rows at the bottom of the table).

The colors of the cells indicate the rate of performance of the four models on each dataset, i.e. row-wise, where the shades of green highlight the most successful results, and reddish shades show the least efficient models.

Table 9. Comparison of one-feature models across datasets (accuracy)

datasets/models	sem	lex	morph	concr
bombardirovat	0.75	**0.76**	0.75	0.63
gret	0.71	**0.86**	0.7	0.8
doit	0.7	**0.85**	0.73	0.81
napadat	0.59	0.74	**0.75**	0.51
ochertit	0.59	0.81	0.77	**0.91**
otrubit	0.64	**0.81**	0.77	0.66
pilit	0.59	**0.8**	0.74	0.71
podkhvatyvat	0.68	0.75	0.76	**0.78**
prichesat	0.69	**0.87**	0.77	0.83
raspylyat	0.79	**0.92**	0.77	0.86
razbavlyat	0.79	**0.93**	0.75	0.71
syedat	0.82	**0.84**	0.74	0.65
trubit	0.78	**0.81**	0.73	0.8
ukolot	0.52	0.73	**0.74**	0.66
utyuzhit	0.65	**0.91**	0.75	0.76
vykraivat	0.81	0.86	0.83	**0.9**
vzorvat	0.51	**0.77**	0.7	0.62
vzvesit	0.5	**0.82**	0.73	0.8
zazhigat	0.6	**0.82**	0.72	0.73
zhonglirovat	0.66	**0.84**	0.7	0.78
combined dataset	0.65	**0.82**	0.67	0.63
mean of 20 datasets	0.67	0.82	0.74	0.74
std of 20 datasets	0.099	0.056	0.03	0.099

Sem – distributional semantic vectors, *lex* – lexical co-occurrence vectors, *morph* – morphosyntactic co-occurrence vectors, *concr* – concreteness vectors

The results of the models on the combined dataset deserve special attention since they allow us to make judgements about the generalizability of the models. The performance on the combined dataset is interesting both in the applied and the theoretical perspective: when a classifier runs on the joint dataset, it receives a mixed signal composed of the signals of all the verbs, and tries to pick up the ones that would allow it to discriminate between the two classes. Whenever the classifier manages to achieve high accuracy, this fact testifies that there exist certain features (e.g. a certain set of lexemes, in the case of the lexical co-occurrence model, or a set of morphological characteristics, in the case of the morphosyntactic co-occurrence model, etc.) which are shared by all or by the majority of the datasets irrespective of the idiosyncratic traits of individual verbs. Consequently, such common features can be expected to serve as reliable predictors of metaphoricity vs. non-metaphoricity, in general. If we reveal the presence of such predictors of metaphoricity, in the practical perspective it may lead to the possibility of building weakly supervised classifiers for metaphor identification; they would require a small annotated dataset for training, after which they may be expected to successfully run on new unseen data. In the theoretical perspective, existence of predictors of metaphoricity might lend empirical support to the theories of metaphor outlined in Section … of Chapter I and in Section 1 of the present chapter.

The mean of each model's performance on 20 datasets shows the average performance of the model across the verbs. It should be compared with the performance of the model on the combined dataset: if the two values are close to each other, it may suggest that the type of feature used to train and test the classifier generalizes well across the datasets.

The standard deviation across the 20 datasets is an indicator of how much the results are spread around the mean. This value can be interpreted as follows: higher standard deviation (the greenish cells in the bottom row of Table 10) means that the performance of the model is inconsistent across the verbs. On the contrary, lower values of standard deviation (the reddish-colored cells) signify consistent and stable performance of the model.

In what follows we will only comment on the performance of the models; as for linguistic interpretation of the patterns which condition such results, it will be provided in the discussion which will be offered in Chapter IV.

It can be seen that *sem* model is overall the least efficient of all: its performance varies in the wide range of 0.5-0.82 (std = 0.99) indicating inconsistent performance of the model across the datasets. For most of the rest of the verbs, the distributional semantic feature delivers the worst or the second worst result. The only verb for which the *sem* model yields a relatively high accuracy is *syedat* 'to eat smth up'. The two verbs with the lowest accuracies are *vzorvat* 'to explode smth' and *vzvesit* 'to weigh smth'. The performance of the model on the combined dataset and the mean of the 20 datasets are close to each other – 0.65 and 0.67, correspondingly; this suggests that the properties of individual datasets are similar to the characteristics of the joint dataset.

The most successful of the four models is the model trained on lexical co-occurrence vectors; its results range between the moderate 0.73 for *ukolot* ‘to prick smb / smth’ and 0.93 for *razbavlyat* ‘to dilute, to liquefy smth’; the latter is the highest of all the results across the datasets and the one-feature models. For an overwhelming majority of the verbs the *lex* model produces either the top or the second best result. The standard deviation is moderate (0.56), indicating a relatively stable performance of the model across the 20 datasets. The accuracy on the combined dataset is relatively high, 0.82; the mean accuracy of the 20 datasets is identically high, 0.82, which means that the lexical co-occurrence feature generalizes well across the target verbs.

The *morph* and the *concr* models both perform with a moderate success: their mean values across the datasets are equal, 0.74. Nonetheless, the two models differ in the standard deviations across the datasets. The standard deviation of the *morph* model is low (0.03), indicating reliable consistency; the standard deviation of the *concr* model is high (0.99), which attests to unstable performance. The accuracy of the *morph* model ranges from 0.7 (*zhonglirovat* ‘to juggle smth’) to 0.83 (*vykraivat* ' to cut (in sewing: parts of a garment, from fabric)’). The best accuracy of the *concr* model – 0.91 is achieved on the dataset of the verb *ochertit* ' to outline smth’, while the lowest result (0.51) is shown on the verb *napadat* ‘to attack smb / smth’.

In total, the *lex* model shows the highest result in the majority of the datasets – 16 of the 21 (these results are highlighted in bold type); the *concr* model comes first on three datasets, and the *morph* model is the best on two datasets. The model based on distributional semantic vectors never beats the other models on individual datasets. This model, *sem*, is the least successful: it achieves the lowest accuracy (the underlined values) on 13 of the 21 datasets. The *concr* model fails on four datasets, and the morph model – on three. The model based on lexical co-occurrence vectors never gives a failing result.

On the combined dataset of 20 verbs, the high achiever is the *lex* model (0.82), while the other three model considerably fall behind and perform at more or less comparable rates: *morph* with the accuracy of 0.67 followed by *sem* with 0.65, and finally, by *concr* with 0.63.

The main conclusion which can be drawn from the above comparison of the one-feature models across the datasets is the following: the most promising model in terms of generalizability is the one based on lexical co-occurrence vectors; its performance is the most successful and stable as compared to the other models. Of the remaining three, the morph model looks somewhat better than *sem* and *concr*: it delivers slightly higher accuracy on the combined dataset than these two, and it behaves in a more consistent fashion (consider the low standard deviation). Between *sem* and *concr*, the former model compares more favorably: it behaves more consistently, although its accuracy is rather moderate. The concreteness-based model, *concr*, is the least stable and efficient of all; however, analysis of this model may lead to valuable implications in terms of linguistic behavior of metaphoric usages.

5.4.2. Comparison of datasets across one-feature models

Table 10. Comparison of datasets across one-feature models (accuracy)

datasets/ models	sem	lex	morph	concr	mean of 4 models	std of 4 models
bombardirovat	0.75	0.76	0.75	0.63	0.72	0.054
gret	0.71	0.86	0.7	0.8	0.77	0.066
doit	0.7	0.85	0.73	0.81	0.77	0.06
napadat	0.59	0.74	0.75	0.51	0.65	0.102
ochertit	0.59	0.81	0.77	0.91	0.77	0.116
otrubit	0.64	0.81	0.77	0.66	0.72	0.072
pilit	0.59	0.8	0.74	0.71	0.71	0.076
podkhvatyvat	0.68	0.75	0.76	0.78	0.74	0.038
prichesat	0.69	0.87	0.77	0.83	0.79	0.068
raspylyat	0.79	0.92	0.77	0.86	0.84	0.059
razbavlyat	0.79	0.93	0.75	0.71	0.8	0.083
syedat	0.82	0.84	0.74	0.65	0.76	0.075
trubit	0.78	0.81	0.73	0.8	0.78	0.031
ukolot	0.52	0.73	0.74	0.66	0.66	0.088
utyuzhit	0.65	0.91	0.75	0.76	0.77	0.093
vykraivat	0.81	0.86	0.83	0.9	0.85	0.034
vzorvat	0.51	0.77	0.7	0.62	0.65	0.097
vzvesit	0.5	0.82	0.73	0.8	0.71	0.127
zazhigat	0.6	0.82	0.72	0.73	0.72	0.078
zhonglirovat	0.66	0.84	0.7	0.78	0.75	0.07
combined dataset	0.65	0.82	0.67	0.63	0.69	0.075

Sem – distributional semantic vectors, *lex* – lexical co-occurrence vectors, *morph* – morphosyntactic co-occurrence vectors, *concr* – concreteness vectors

The central part of Table 11 replicates the information from the central part of Table 10, but this time the color scheme is adjusted vertically, in the column-wise direction – the greener cells in the columns indicate the datasets which yield the best performance with a given model, while redder shades color the datasets that failed with this particular model. The two rightmost columns feature the mean and the standard deviation of the four models across each dataset. Standard deviation illustrates the consistency in the performance of the four models on individual verbs. Lower values (the reddish cells) imply that all the models perform equally well or equally poorly. Higher standard deviation (the cells filled with green) mean that some of the models substantially outperform the others.

The following paragraphs of this subsection will only comment on how datasets differ in their performance with each of the four models; an in-depth linguistic analysis of such behavior will offered in Chapter IV.

The classifier trained on the distributional semantic vectors proved the most successful on the verbs *syedat* ‘to eat smth up’ (0.82), *vykraivat* ' to cut (in sewing: parts of a garment, from fabric)’ – 0.81, and *raspylyat* ‘to spray smth’ and *razbavlyat* ‘to dilute, to liquefy smth’ – both 0.79. The lowest accuracy – just about the chance level – was obtained with the verbs *vzorat* ‘to explode smth’ (0.5), *vzvesit* ‘to weigh smth’ (0.51), and *ukolot* ‘to prick smb / smth’ (0.52).

The model trained with the lexical co-occurrence vectors was the most accurate on the datasets of *razbavlyat* ‘to dilute, to liquefy smth’ (0.93), *raspylyat* ‘to spray smth’ (0.92), and *utyuzhit* ‘to iron smth’ (0.91). The following verbs fell behind the others: ukolot ‘to prick smb / smth’ (0.73), napadat ‘to attack smb / smth’ (0.74), *podkhvatyvat* ‘to catch (smth falling)’ – 0.75, and bombardirovat ‘to bombard smb / smth’ (0.76).

The lowest performance of the morphosyntactic classifier was achieved on the combined dataset of 20 verbs (0.67), which was followed by *zhonglirovat* ‘to juggle smth’ and *gret* ‘to heat / warm smb / smth’ with 0.7, and then *zazhigat* ‘to ignite smth’ with 0.72. The highest accuracy was recorded on the verb *vykraivat* ' to cut (in sewing: parts of a garment, from fabric)’ with 0.83, which in terms of accuracy stands far apart from the other datasets when running on morphosyntactic co-occurrence vectors.

The leaders on the performance with the concreteness-based model are the datasets ochertit ‘to outline smth’ (0.91) and *vykraivat* ' to cut (in sewing: parts of a garment, from fabric)’ (0.9), followed by *raspylyat* ‘to spray smth’ with 0.86. The absolute failure of the *concr* model – the accuracy around the chance level – occurs on the dataset of *napadat* ‘to attack smb / smth’ (0.5).

In total, the most successful datasets in the aggregate terms of the four models (see the column *‘mean of 4 models’*) are the verbs: *vykraivat* ' to cut (in sewing: parts of a garment, from fabric)’ – *mean* = 0.85, *raspylyat* ‘to spray smth’ (0.84), *razbavlyat* ‘to dilute, to liquefy smth’ (0.8), and *prichesat* ‘to comb smb / smth’ (0.79). The lowest mean accuracy across the four models was achieved on the datasets of *vzorvat* ‘to explode smb / smth’ and *napadat* ‘to attack smb / smth’ – both 0.65, as well as on *ukolot* ‘to prick smb / smth’ with 0.66.

When analyzing the consistence in the performance of the four models on each individual dataset (the rightmost column, ‘*std of dataset*’), it is worth noting the low standard deviation on the verbs *trubit* ‘to trumpet’ with *std* = 0.031, *vykraivat* ' to cut (in sewing: parts of a garment, from fabric)’ with std = 0.034, and *podkhvatyvat* ‘to catch (smth falling)’ with *std* = 0.038. All the four models on these datasets yield result that comfortably fit within a relatively narrow range: for *trubit*, the best model is *lex* (0.81), the second best is *concr* (0.8) which is closely followed by *sem* (0.78), and finally, by *morph* (0.73). In the case of *vykraivat*, the highest accuracy is yielded by the *concr* model (0.9), the runner-up is *lex* with 0.86 followed by *morph* and *sem* in close succession (0.83 and 0.81, respectively). The dataset of *podkhvatyvat* is similar to the previous verb in that the best result is achieve with the *concr* model (0.78), while *morph* and *lex* perform with 0.76 and 0.75, correspondingly; lastly, the lowest result is demonstrated by the *sem* model.

Four of the datasets have high values of standard deviation when comparing the four models on individual datasets: *vzvesit* ‘to weigh smth’ (*std* = 0.127), *ochertit* ‘to outline smth’ (*std* = 0.116), *napadat* ‘to attack smb / smth’ (*std* = 0.102), and *vzorvat* ‘to explode smth’ (*std* = 0.097). Such values of standard deviation speak of sharp contrasts among the results of the models, when some of them achieve high results while the others are a failure. For *vzvesit*, the best result is 0.82 with the *lex* model, the lowest is 0.5 with the *sem* model; *the* concr model is close to the leader with 0.8, and *morph* lingers in the middle of the range with 0.73. *Ochertit* boasts of one of the highest results among all the models and the datasets – 0.91 with the *concr* model; however, it sinks to 0.59 with the sem model, while *lex* and *morph* are moderately successful with 0.81 and 0.77, respectively. In the case of *napadat*, the two outsiders are the models *concr* and *sem* (0.51 and 0.59), while *morph* and *lex* run next to each other with 0.75 and 0.74, correspondingly. The accuracies on the dataset of *vzorvat* fluctuate between the failing 0.51 (*sem*) and the moderately strong 0.77 (*lex*), while *lex* and *morph* occupy the medium range (0.77 and 0.7, respectively).

To sum up, some of the verbs are successfully classified with all or at least with three of the models, while others perform more or less well with one or two models and fail with the other models. Moreover, each of the models performs better on some datasets and fall short on the others. Apparently, this primarily has to do with the patterns of lexical and morphosyntactic combinability of the meanings of target verbs, and the ensuing semantic and concreteness characteristics of their arguments. The next chapter of the thesis will look at the patterns in the behavior of individual verbs which lead to efficiency of one models and inefficient performance of the others.

5.4.3. Evaluating the complexity of models

Beside the one-feature models whose results were presented and analyzed in the previous section, we also experimented with multi-feature classifiers, i.e. with classifiers that were run on combinations of one, two, three, and four features.

Table 12 summarizes the accuracy of all the 15 models; the coloring of the cells changes along the rows, i.e. the high results on each dataset are colored with green and greenish while the low performance is indicated with red. The rest of this subsection will analyze whether increasing complexity of models leads to higher accuracy of classification, and if it does – whether the utility of the gain offsets the complexity of the model.

The cells of Table 12 contain additional notation to indicate the three best results on each dataset. Exclamation mark ("!") signals the top result(s) on a dataset; asterisk ("*") points at the second best value(s), and hashtag ("#") indicates the third best accuracy. The second and third results are indicated as being such if they fall within the range of one and two percentage points from the top result, correspondingly. The first, second, and third best results will be collectively referred to as 'optimal' results in the next paragraphs.

We can see that on ten of the 20 individual verbs datasets – as well as on the combined dataset – an optimal result is achieved already with one of the least complex, one-feature models (the datasets of the verbs *gret*, *doit*, *napadat*, *ochertit*, *pilit*, *prichesat*, *raspylyat*, *razbavlyat*, *utyuzhit*, *zhonglirovat*, and the combined dataset). Among the remaining ten datasets, eight cases require a model of increased complexity – a two-feature model (the verbs *bombardirovat*, *syedat*, *trubit*, *ukolot*, *vykraivat*, *vzorvat*, *vzvesit*, *and zazhigat*). And only on two datasets (*otrubit* and *podkhvatyvat*) to achieve an optimal accuracy the combination of three features has to be applied. On the combined dataset, the top result is obtained with two-, three-, and four-feature combinations; however, the second best result becomes possible already with a one-feature model (the *lex* model) – as we pointed out above (Section 4.1.1), this model seems to hold the maximum potential for generalizability of all the tested one-feature models.

In order to asses the utility of the gain provided by increasing the complexity of the model we used Formula (7), where: U is the utility of the gain in accuracy, IR is the inverse rank of a result achieved by a model (i.e. $Rank - (Rank - 1)$), $N_{features}$ is the number of features in the model, and $Max_{dataset}$ is the maximum accuracy achieved on a given dataset; thus, utility equals the inverse rank of a result, which is weighted by the number of features and normalized by the maximum accuracy. The maximum possible value of U is 1, and the possible minimum is 0.1.

$$U = \left(\frac{IR}{N_{features}}\right) \Big/ Max_{dataset} \tag{7}$$

Table 11. Accuracy and complexity of models

	one-feature models				two-feature models						three-feature models and the four-feature model				
datasets/models	**sem**	**lex**	**morp**	**concr**	**sem+lex**	**sem+ morph**	**sem+ concr**	**lex+ morph**	**lex+concr**	**morph+ concr**	**sem+lex+ morph**	**sem+lex+concr**	**sem+morph+ concr**	**lex+morph+ concr**	**sem+lex+ morph+concr**
bombardirovat	0.75	0.76	0.75	0.63	***0.81**	0.77	0.75	0.79	0.72	0.76	***0.81**	**! 0.82**	0.78	***0.81**	**#0.8**
gret	0.71	***0.86**	0.7	0.8	***0.86**	0.74	0.8	**#0.85**	***0.86**	0.8	**! 0.87**	***0.86**	0.8	***0.86**	**!0.87**
doit	0.7	**#0.85**	0.73	0.81	0.75	***0.86**	**! 0.87**	0.68	0.74	**#0.85**	0.78	0.81	0.78	***0.86**	0.84
napadat	0.59	***0.74**	**! 0.75**	0.51	***0.74**	**#0.73**	0.59	0.72	***0.74**	**! 0.75**	0.7	***0.74**	**#0.73**	**! 0.75**	**#0.73**
ochertit	0.59	0.81	0.77	**! 0.91**	0.79	0.8	**! 0.91**	0.81	0.82	**! 0.91**	0.8	0.81	***0.9**	0.82	0.82
otrubit	0.64	0.81	0.77	0.66	0.8	0.77	0.67	0.81	0.81	0.79	***0.85**	0.79	0.78	0.8	**! 0.86**
pilit	0.59	***0.8**	0.74	0.71	0.74	0.74	0.73	**#0.79**	**#0.79**	0.76	0.74	0.72	0.77	**! 0.81**	0.74
podkhvatyvat	0.68	0.75	0.76	0.78	0.79	0.76	0.81	0.74	0.74	0.81	***0.84**	**#0.83**	0.82	0.75	**! 0.85**
prichesat	0.69	**! 0.87**	0.77	0.83	0.84	0.77	***0.86**	0.8	0.77	**! 0.87**	0.73	0.79	**! 0.87**	0.76	0.76
raspylyat	0.79	***0.92**	0.77	0.86	**#0.91**	0.82	0.87	**! 0.93**	***0.92**	0.9	***0.92**	**#0.91**	0.89	***0.92**	**! 0.93**
razbavlyat	0.79	**! 0.93**	0.75	0.71	**#0.91**	0.81	0.79	**#0.91**	**! 0.93**	0.79	***0.92**	**! 0.93**	0.81	***0.92**	**#0.91**
syedat	0.82	0.84	0.74	0.65	**#0.86**	0.82	0.84	**! 0.88**	0.84	0.77	**! 0.88**	***0.87**	0.84	***0.87**	**#0.86**
trubit	0.78	0.81	0.73	0.8	0.78	0.8	***0.83**	0.76	0.73	**#0.82**	0.79	0.72	**! 0.84**	0.76	0.77
ukolot	0.52	0.73	0.74	0.66	0.72	**#0.76**	0.66	0.75	0.75	***0.77**	***0.77**	0.72	**! 0.78**	**#0.76**	0.75
utyuzhit	0.65	**! 0.91**	0.75	0.76	0.87	0.77	0.76	***0.9**	**#0.89**	0.8	**#0.89**	**! 0.91**	0.81	***0.9**	***0.9**
vykraivat	0.81	0.86	0.83	**#0.9**	0.85	0.86	**#0.9**	0.87	0.83	**! 0.92**	0.88	0.86	***0.91**	0.88	0.85
vzorvat	0.51	0.77	0.7	0.62	0.73	0.7	0.6	0.72	**! 0.8**	0.72	0.67	0.72	0.71	0.73	0.69
vzvesit	0.5	0.82	0.73	0.8	0.8	0.74	0.79	***0.85**	**! 0.86**	**! 0.86**	***0.84**	***0.85**	**! 0.86**	0.77	0.76
zazhigat	0.6	0.82	0.72	0.73	0.78	0.72	0.75	0.79	0.81	0.77	0.77	**! 0.85**	0.79	0.8	0.81
zhonglirovat	0.66	**#0.84**	0.7	0.78	0.73	0.74	0.79	**! 0.86**	0.72	0.78	0.77	0.71	0.78	0.74	0.77
combined dataset	0.65	***0.82**	0.67	0.63	***0.82**	0.71	0.68	**! 0.83**	**! 0.83**	0.71	**! 0.83**	**! 0.83**	0.73	**! 0.83**	**! 0.83**

! – the first best result on the dataset; * – the second best result, # – the third best result

Table 12. Normalized weighted gain of model complexity

	one-feature models				two-feature models						three-feature models and the four-feature model				
datasets	**sem**	**lex**	**morp**	**concr**	**sem+lex**	**sem+morph**	**sem+concr**	**lex+pos**	**lex+concr**	**pos+concr**	**sem+lex+morph**	**sem+lex+concr**	**sem+morph +concr**	**lex+morph +concr**	**sem+lex +morph+concr**
bombardirovat	0.67	0.89	0.67	0.22	**1**	0.56	0.33	0.78	0.22	0.44	0.67	0.74	0.44	0.67	0.44
gret	0.33	**1**	0.17	0.67	0.5	0.25	0.33	0.42	0.5	0.33	0.39	0.33	0.22	0.33	0.29
doit	0.22	**1**	0.33	0.78	0.28	0.56	0.61	0.06	0.22	0.5	0.22	0.26	0.22	0.37	0.22
napadat	0.29	0.86	**1**	0.14	0.43	0.36	0.14	0.29	0.43	0.5	0.14	0.29	0.24	0.33	0.18
ochertit	0.13	0.63	0.25	**1**	0.19	0.25	0.5	0.31	0.38	0.5	0.17	0.21	0.29	0.25	0.19
otrubit	0.13	**1**	0.5	0.25	0.44	0.25	0.19	0.5	0.5	0.38	0.38	0.25	0.21	0.29	0.31
pilit	0.11	**1**	0.56	0.22	0.28	0.28	0.22	0.44	0.44	0.33	0.19	0.11	0.26	0.37	0.14
podkhvatyvat	0.2	0.6	0.8	**1**	0.6	0.4	0.7	0.2	0.2	0.7	0.67	0.6	0.53	0.2	0.55
prichesat	0.1	**1**	0.4	0.7	0.4	0.2	0.45	0.3	0.2	0.5	0.07	0.17	0.33	0.1	0.08
raspylyat	0.22	**1**	0.11	0.44	0.44	0.17	0.28	0.56	0.5	0.39	0.33	0.3	0.22	0.33	0.28
razbavlyat	0.43	**1**	0.29	0.14	0.36	0.29	0.21	0.36	0.5	0.21	0.29	0.33	0.19	0.29	0.18
syedat	0.8	**1**	0.4	0.2	0.6	0.4	0.5	0.8	0.5	0.3	0.53	0.47	0.33	0.47	0.3
trubit	0.63	**1**	0.25	0.88	0.31	0.44	0.63	0.19	0.13	0.56	0.25	0.04	0.46	0.13	0.13
ukolot	0.2	0.8	**1**	0.4	0.3	0.7	0.2	0.6	0.6	0.8	0.53	0.2	0.6	0.47	0.3
utyuzhit	0.1	1	0.2	0.3	0.35	0.2	0.15	0.45	0.4	0.25	0.27	0.33	0.2	0.3	0.23
vykraivat	0.14	0.57	0.29	**1**	0.21	0.29	0.5	0.36	0.14	0.64	0.29	0.19	0.38	0.29	0.11
vzorvat	0.1	**1**	0.6	0.3	0.45	0.3	0.1	0.4	0.55	0.4	0.13	0.27	0.23	0.3	0.13
vzvesit	0.13	**1**	0.25	0.88	0.44	0.19	0.38	0.63	0.69	0.69	0.38	0.42	0.46	0.21	0.13
zazhigat	0.1	**1**	0.2	0.3	0.3	0.1	0.2	0.35	0.45	0.25	0.17	0.37	0.23	0.27	0.23
zhonglirovat	0.1	**1**	0.2	0.8	0.25	0.3	0.45	0.55	0.2	0.4	0.23	0.1	0.27	0.2	0.18
combined dataset	0.29	**1**	0.43	0.14	0.5	0.36	0.29	0.57	0.57	0.36	0.38	0.38	0.29	0.38	0.29

Table 9 presents the normalized weighted gains of model complexity computed according to Formula (7). As we can see, it is only on one of the datasets (*bombardirovat*) that the maximum utility is obtained with a two-feature model (*sem + lex*). Apart from it, almost all of the maximum utilities are located in the one-feature models area, including the utility on the combined dataset: its maximum utility once again confirms that the *lex* model is likely to be the optimal choice for generalizing the metaphor identification classifier and applying it to new unseen data.

To conclude the section on the results of the metaphor classification experiment, we would like to emphasize that the robustness of lexical co-occurrence features demonstrated by the experiment and the analysis of its results is consonant with the findings of Klebanov, Leong, Heilman, and Flor (2014). They used lexical unigrams (without stemming or lemmatization) to train a logistic regression classifier, among the other features (such as part-of-speech tags, concreteness, and topic vectors). The experiment was designed in the token-level paradigm, where each word in running text had to be classified as either a metaphor or non-metaphor. The classifier was run on several English-language datasets:

1) A set of graduate student essays on topic A;
2) A set of graduate student essays on topic B;
3) A sample from the News subcorpus of the VU Amsterdam Metaphor Corpus (VUAMC, see Section … of Chapter I for more detail);
4) A sample from the Conversation subcorpus of VUAMC;
5) A sample from the Fiction subcorpus of VUAMC;
6) A sample from the Academic subcorpus of VUAMC.

The researchers report that 'simple unigram baseline that achieves surprisingly good results for some of the datasets'; the best results were obtained in the setting when the classifier was trained and tested on the same set of student essays. The performance declined but still remained substantial when the classifier was trained on one set of essays and tested on the other; the authors note that about half of metaphors was shared by the two essay datasets, which is likely to have ensured that the model generalized well across them. Results comparable to the across-datasets essay setting were achieved on the News partition of VUAMC. As for the other partitions of VUAMC (Conversation, Fiction, and Academic), the results were much lower. The authors account for it by the fact that the News dataset tends 'to discuss a set of related topics and exhibit substantial sharing of metaphors across texts'; moreover, the News partition has a larger number of short texts – as opposed to the other three partitions which have small numbers of long texts and discuss highly divergent topics.

The datasets in the experiment of Klebanov et al. were designed according to topics and domains; our experimental dataset is centered around target verbs. Nevertheless, such verb-centric design also entails topical specialization of datasets, since each target verb and its meanings evoke certain sets of topics associated with them, e.g. the central literal meaning of the verb *doit* 'to milk' naturally attracts the topics of farming, dairy production, rural life, etc.; the central literal meaning of the verb *vykraivat* ' to cut (in sewing: parts of a garment, from fabric)' associates with the topics of clothing design and manufacturing, and so on. Thus, in an indirect way, our dataset may be said to bear resemblance to the dataset of Klebanov et al. in this respect. The difference is in the size of contexts comprising the datasets: our dataset is composed of isolated sentences, while the Klebanov et al. datasets are made up of whole texts (e.g. entire essays, in the first two cases, or news bulletins, in the second). This might be one of the factors due to which in our experiment the lexical co-occurrence feature generalized so well across the datasets of individual target verbs (recall the 0.82 accuracy of *lex* on the combined dataset). On top of it, we preprocessed the input text: we used lemmatized word forms and composed our vectors of metaphor association scores (see Section 3.2.1) rather than raw strings.

Generalizabiltiy of the lexical features means that there are certain words which serve as predictors of metaphoricity, on the one hand, and as predictors of non-metaphoric discourse, on the other. In Section … of Chapter IV we will attempt to collect such predictors of metaphoricity using the metaphor association measure adopted in the current experiment. Moreover, we rely the predictions from the theory of embodied metaphor (see Section 1.4) and hypothesize that abstract words will predominate among the top metaphor-associated words and, conversely, the most non-metaphor associated words will be predominantly concrete. This hypothesis will be tested in Section … of Chapter IV.

Summary of Chapter III

In this chapter we explained the theoretical motivation which prompted the choice of the features that were implemented in the metaphor classification experiment. The choice of distributional semantic features was motivated by componential models of word meaning and the selectional preference violation view of metaphor (Katz & Fodor, 1963; Wilks, 1978), as well as the theory of conceptual metaphor (Lakoff & Johnson, 1980a, 1980b). Moreover, the choice of distributional semantic features was promoted by the models of semantic combinability developed by the Moscow Semantic School (Apresyan, 1995, 2009), and the idea of associative chains (Halliday &

Hasan, 1976). The motivation behind the lexical co-occurrence features ensues from the same theories that motivated the distributional semantic features, since the conceptual and semantic schemata realize themselves on the surface level as their lexical manifestations in the text. Morphosyntactic co-occurrence features rely on the premises of the integrative model of lexical semantics developed by the Moscow Semantic School (Apresyan, 1995, 2009) and Construction Grammar (Brugman, 1988; Fillmore, 1985, 1988; Fillmore et al., 1988; Goldberg, 1995; Lakoff, 2008; Lambrecht, 1994). The concreteness feature is based on the tents of the theory of embodied and grounded cognition, and the notion of primary metaphor (Barsalou, 2008, 2010; Lakoff & Johnson, 1999). The feature which is based on occurrences of lexical signals of metaphoricity ('flag words') and quotation marks was promoted by the body of previous scholarship on metaphor and the practice of annotating metaphor corpora (Goatly, 1997; G. Steen et al., 2010).

Further in the chapter we laid out the technical details of feature engineering, i.e. the computational models and the statistical measures which were used in order to vectorize the textual data from the experimental corpus. The distributional semantic vectors 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. The lexical co-occurrence vectors are computed by means of the ΔP metric (Levshina, 2015); the obtained scores show the metaphor association indexes of lexemes in the corpus. The 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). The indexes of concreteness are computed using the seed list of approx. 500 concrete ('thingness') 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. The flag words and quotation vectors are computed as the number of their occurrences and the distance 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.

Then, we reported the results of the metaphor classification experiment and compared the performance of the single-feature models. 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 to be 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 will 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 will also be conducted in Chapter IV.

Appendix 1. Annotator guidelines for the inter-annotator reliability test (Chapter II. Section 3.2)

ИНСТРУКЦИЯ АННОТАТОРА

Уважаемый аннотатор,

Благодарим Вас за согласие принять участие в разметке данных для нашего проекта. Проект направлен на выявление представлений о критериях метафоричности глаголов у носителей русского языка. В работе принимают участие несколько аннотаторов, каждый их которых произведет свою разметку; после этого мы вычислим меру согласованности между решениями, принятыми аннотаторами. Полученные данные будут использованы в экспериментах с использованием машинного обучения, цель которых – научиться автоматически распознавать метафорические употребления глаголов.

1

Перейдя по полученной Вами ссылке, Вы попадаете в таблицу Google Spreadsheet, которая состоит из 20 листов. Название каждого листа соответствует одному из 20 глаголов, с которыми Вам предстоит работать; число после нижнего подчеркивания обозначает количество предложений на листе (Рис. 1).

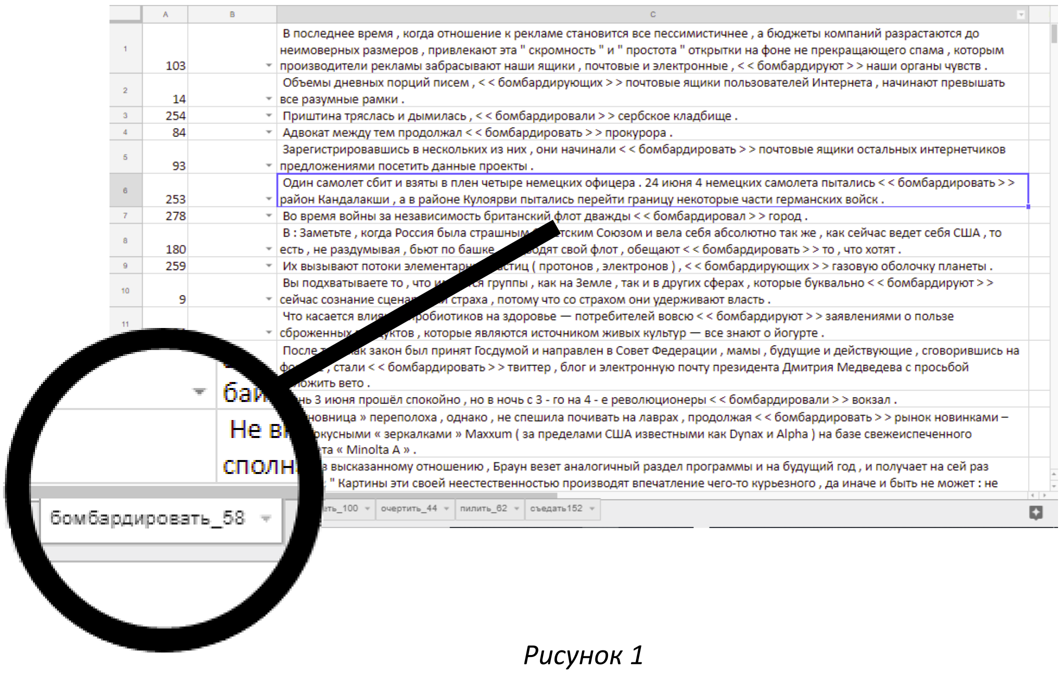


Рисунок 1

Количество предложений на листах варьируется от 45 до 160; в общей сложности вам предстоит разметить около 1450 предложений.

2

Предложения, подлежащие разметке, находятся в колонке «Текст предложения». Каждое предложение содержит тот глагол, который вынесен в заголовок листа (т.е. каждое предложение на листе с заголовком «бомбардировать» содержит глагол «бомбардировать»). В тексте каждого предложения целевой глагол с двух сторон выделен двойными треугольными скобками (Рис. 2).

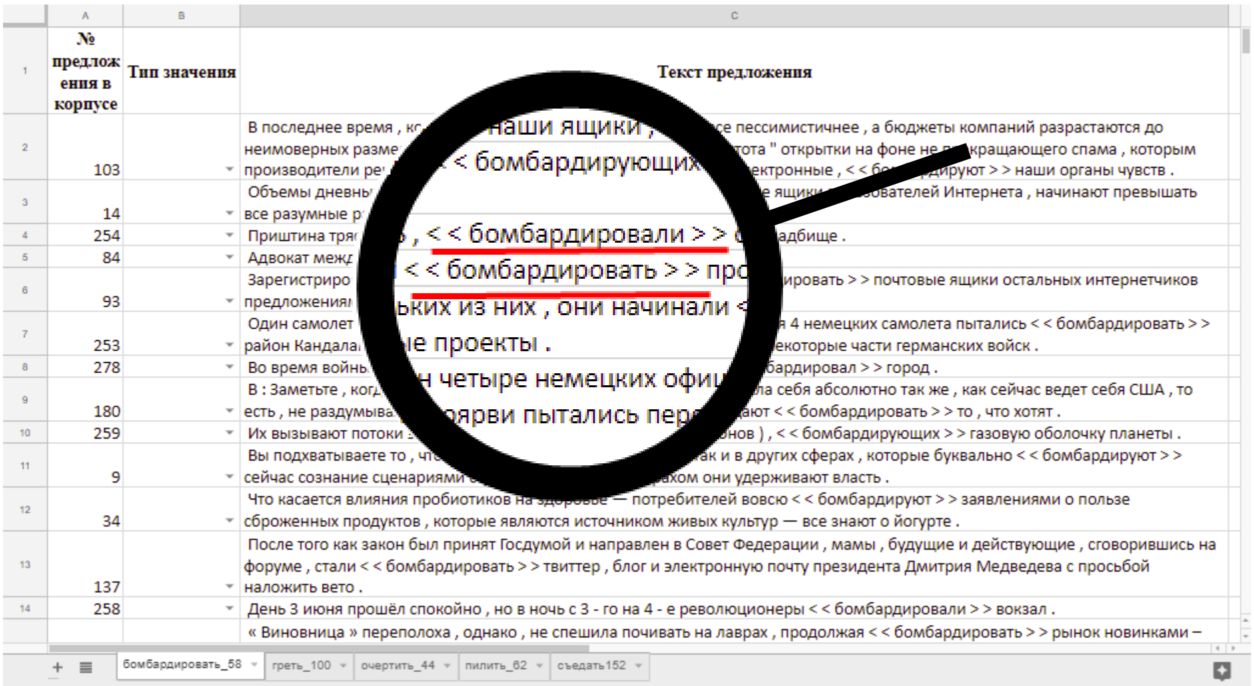


Рисунок 2

> **Предупреждение**: предложения взяты из интернета, вследствие чего могут содержать ненормативную лексику или контент пограничного содержания; заранее приносим свои извинения.

Вы можете вносить следующие изменения в документ: изменять шрифт текста, ширину колонок и высоту строк.
Мы просим Вас **не вносить в документ следующих изменений**: удалять столбцы или строки; менять порядок столбцов или строк; **каким-либо образом редактировать содержимое столбцов и строк, вносить изменения в содержащийся в них текст**.

3

Порядок работы

Ваша основная задача – отобрать все предложения, в которых глагол употребляется **<u>В ПРЯМОМ (т.е. ИСХОДНОМ, НЕМЕТАФОРИЧЕСКОМ значении)</u>**.

1. Для того, чтобы определить для себя прямое значение глагола, представьте себе **<u>наиболее прототипическое действие, которое может обозначаться данным глаголом</u>**.

Например, для глагола «*состряпать*» прототипическое исходное прямое значение можно проиллюстрировать следующим предложением:

Мартин < < состряпает > > тыквенный пирог , а Хина приготовит сочный бифштекс.

2. Прочитайте одно предложение, уделяя особое внимание выделенном глаголу. Если значение глагола в данном контексте соответствует Вашему представлению о его прямом значении, то в колонке «Тип значения» в выпадающем списке выберите пункт «ПРЯМОЕ ЗНАЧЕНИЕ»; ячейка окрасится в зеленый цвет.
3. Если значение глагола в рассматриваемом контексте не соответствует Вашему представлению о прямом значении, имеющемся у данного глагола, то в колонке «Тип значения» выберите пункт «ПЕРЕНОСНОЕ ЗНАЧЕНИЕ»; ячейка окрасится в оранжевый цвет.

Например, в следующем предложении глагол «состряпать» употреблен в переносном значении:

Вместе они быстро < < состряпали > > текст письма.

ОБРАТИТЕ ВНИМАНИЕ, что к типу «ПЕРЕНОСНОЕ ЗНАЧЕНИЕ» следует относить как непосредственно переносные значения глагола, так и его **<u>употребления в качестве необычных, художественных метафор</u>**, а также в составе **<u>устойчивых идиоматических выражений</u>**.

4. Если рассматриваемое Вами употребление глагола не полностью соответствует Вашему представлению о прототипическом прямом значении, но при этом Вы не можете однозначное отнести его к переносным значениям, выберите пункт «СКОРЕЕ ПРЯМОЕ»; ячейка окрасится в желтый цвет.
5. Если рассматриваемое Вами употребление глагола полностью не соответствует Вашему представлению о прототипическом прямом значении, и в то же время Вы не можете однозначно отнести его к переносным значениям, выберите пункт «СКОРЕЕ ПЕРЕНОСНОЕ»; ячейка окрасится в голубой цвет.
6. Если после прочтения предложения Вы не смогли определить, в каком значении в нем употребляется глагол, выберите пункт «НЕПОНЯТЕН СМЫСЛ» выпадающего меню; ячейка окрасится в красный цвет.

Используйте этот пункт в следующих случаях: предложение слишком короткое; в предложении используются непонятные слова; предложение содержит структурные или содержательные ошибки, затрудняющие понимание. Пожалуйста, постарайтесь **по минимуму** использовать пункт «НЕПОНЯТЕН СМЫСЛ» таким образом, чтобы их количество на листе составляло не более 2-4 раз.

Пожалуйста, при принятии всех решений **руководствуйтесь только собственными соображениями**, не советуясь ни с кем.

Вы можете **в любое время вернуться к любой из сделанной вами разметок и изменить принятое Вами решение**. Для того, чтобы изменить тип значения, выберите другой пункт выпадающего меню. Удалить значения в ячейке «Тип значения» можно нажатием клавиши Delete или Backspace.

При желании Вы можете оставить свои комментарии к сделанным Вами аннотациям; для этого воспользуйтесь колонкой «Комментарий аннотатора (при необходимости)» (справа от колонки «Текст предложения». Пожалуйста, обратите внимание, что не нужно комментировать каждое принятое Вами решение.

7. После разметки всех находящихся на листе предложений, пожалуйста, просмотрите весь лист еще раз: на нем не должно остаться незаполненных белых ячеек в колонке «Тип значения» (в каждой строке должно быть выставлено одно из значений выпадающего списка).
8. По завершении работы с одним глаголом переходите к предложениям со следующим глаголом, расположенным на следующем листе. Порядок работы с листами не имеет значения: Вы можете размечать глаголы в любом порядке.

Желаем Вам успешной работы!

Appendix 2. Concrete ('thingness') paradigm words (Chapter III. Section 3.4)

авиабомба
авоська
автомагнитола
авторучка
акваланг
аккордеон
аккумулятор
аксельбант
аптечка
арматура
арфа
афиша
баллон
бандаж
бандана
бандероль
банджо
бант
барабан
бас-гитара
батарейка
баул
бахила
бачок
башмак
баян
безрукавка
бейсболка
белье
бензонасос
бензопила
берет
бечевка
бигуди
бидон
билет
бинокль
бинт
бирка
бланк
блендер
блесна
блокнот
блуза
блюдце
блестка
бобина
бокал
болт
босоножка
ботинок
ботфорт
бочка
бочонок
браслет
брелок
бретелька
бритва
бронежилет
брошь
брошюра
брюки
будильник
булавка
булыжник
бумажник
бусы
бутылка
бушлат
бюстгальтер
ваза
валенки
валторна
варежка
вафельница
веб-камера
ведро
веер
веник
веретено
вертел
веревка
ветровка
видеокамера
видеокарта
видеокассета
видеомагнитофон
вилка
вилы
винтовка
виолончель
водолазка
водонагреватель
вольтметр
воротник
втулка
выкройка
вышивка
гаджет
газонокосилка
галоша
галстук
гантели
гармошка
гарпун
гвоздик
гвоздодер
гвоздь
гимнастерка
гирлянда
гиря
гитара
гобой
горшок
готовальня
грабли
градусник
граммофон
граната
гранатомет
графин
гребенка
грелка
двустволка
девайс
джемпер
джинсы
джойстик
диадема
диктофон
дискета
дождевик
дозиметр
домкрат
дрель
дробовик
дубленка
дуло
дуршлаг
душегрейка
дырокол
жакет
жгут
жезл
железяка
жетон
жилет
зажим
заколка
замочек
записка
запонка
зачетка
защелка
зеркальце
зипун
зонт
зубило
зубочистка
игрушка
кадило
кайло
календарик
калькулятор
кальсоны
канат
кандалы
канделябр
канистра
капельница
капюшон
карандаш
картридж
картуз
каска
кассета
кастрюля
кашпо
кегля
кепка
кимоно
кинопленка
кинопроектор
кипятильник
кирка
кисет
кисточка
китель
клавиатура
клавиша
кларнет
клатч
клемма
клюшка
кляп
кобура
ковбойка
колготки
колечко
колчан
колье
кольчуга
комбинезон
компостер
компьютер
конденсатор
контрабас
конфетница

конфорка
корзина
коробок
коробочка
коромысло
корсет
кортик
косметичка
костыль
костюм
косынка
котелок
котомка
кофеварка
кофейник
кофемашина
кофемолка
кофта
кошелек
кроссовки
крюк
крючок
ксерокс
ксилофон
кувшин
кулер
кулон
куртка
кусачки
кушак
лампада
лампочка
ластик
ледоруб
лейка
лейкопластырь
леска
линейка
листовка
лобзик
ложка
лопата
лукошко
лыжа
магнитофон
майка
мандолина
манжет
медаль
медальон
металлодетектор
металлоискатель
мешок
мигалка
микрокалькулятор
микропроцессор
микроскоп
микрофон
миксер
миномет
миска
мобильник
модем
мокасин
мольберт
монитор
монтировка
мотыга
мочалка
мультиварка
мундштук
мухобойка
мыльница
мясорубка
мяч
наволочка
нагайка
наган
наперсток
насос
наушники
нож
ножницы
ножовка
носок
ноутбук
ночнушка
обогреватель
ободок
обруч
обувь
огнемет
огнетушитель
одеяло
ожерелье
окуляр
окурок
олимпийка
отвертка
открытка
очки
ошейник
палка
пальто
памперс
панама
папироска
парик
пароварка
парогенератор
паяльник
педаль
пеленка
пенал
пепельница
перстень
перфоратор
перчатка
пиала
пиджак
пижама
пинцет
пипетка
пистолет
платок
платье
плащ
плуг
плетка
поварешка
погон
погремушка
подгузник
подкова
подсвечник
подстаканник
подтяжки
подшлемник
подштанники
полотенце
полуботинок
полупальто
полушубок
портсигар
портупея
портфель
портьера
портянка
посох
принтер
присоска
прихватка
прищепка
пробирка
промокашка
простыня
протез
противень
пружинка
пряжа
пряжка
пуговица
пудреница
пулемет
пуловер
пульверизатор
пуховик
пылесос
ранец
расческа
револьвер
резиночка
ремень
рогатка
рояль
рубанок
рубашка
ружье
рукав
рукавица
рукомойник
рукоятка
рюкзак
ряса
сабля
саквояж
саксофон
салатник
салатница
салфетка
самовар
сандалия
сапог
сарафан
сверло
светильник
светодиод
свистулька
свитер
секундомер
селедочница
сережка
сигара
сигаретка
скакалка
скалка
скальпель
скатерть
скафандр
скипетр
сковорода
скороварка
скребок
скрипка
смартфон
соковыжималка
солонка
сомбреро
сорочка
спецовка
спиннинг
спичка
стакан
стамеска
статуэтка
стеклорез
степлер
стремя

сумка
сумочка
сундучок
супница
сюртук
тапок
тарелка
телевизор
телекамера
тетрадь
топор
тостер
точилка
тромбон
трость
трусы
туфля
тюбетейка
тюбик
тяпка
удочка
узелок
указка
уключина
утюг
ушанка
фартук
фитиль
флейта
фломастер
флэшка
фляга
фортепьяно
фотоаппарат
фотообъектив
фрак
фуганок
фужер
фуражка
футболка
футляр
халат
хлебница
хлястик
циркуль
циферблат
чайник
чашка
чемодан
чепчик
чернильница
чулок
шаль
шапка
шарф
швабра
шило
шинель
шкатулка
шланг
шлем
шляпа
шлепанец
шпага
шпингалет
шприц
шпулька
штаны
штык-нож
шуба
шуруп
щеколда
щетка
электробритва
электрогитара
электродрель
электропила
электрочайник
эспандер
юбка

List of References

Ahmad, K., Gillam, L., & Tostevin, L. (2000). Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, 1–8.

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147–149.

Apresyan, Y. (1995). *Leksicheskaya semantika. Sinonimicheskiye sredstva yazyka [Lexical semantics. The synonymic means of the language]* (2nd ed.). Moscow: Yazy`ki slavyanskoj kul`tury` (Languages of Russian Culture Publishing House).

Apresyan, Y. (2009). *Issledovanija po semantike i leksikografii: Paradigmatika. [The study of semantics and lexicography: The paradigmatic aspect.]* (Vol. 1). Moscow: Yazy`ki slavyanskoj kul`tury` (Languages of Russian Culture Publishing House).

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, *59*, 617–645.

Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Topics in Cognitive Science*, *2*(4), 716–724.

Benko, V., & Zakharov, V. (2016). Very large Russian corpora: new opportunities and new challenges. In *Computational linguistics and intellectual technologies*. Russian State University for the Humanities.

Birke, J., & Sarkar, A. (2006). A clustering approach for nearly unsupervised recognition of nonliteral language. *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Boguslavsky, I. (2014). SynTagRus–a Deeply Annotated Corpus of Russian. *Les Émotions Dans Le Discours-Emotions in Discourse*, 367–380.

Brugman, C. (1988). *The syntax and semantics of HAVE and its complements*.

Calvo, P., & Gomila, T. (2008). *Handbook of cognitive science: An embodied approach*. Elsevier.

Cameron, L. (2003). *Metaphor in educational discourse*. Retrieved from https://www.google.com/books?hl=en&lr=&id=m4ixAwAAQBAJ&oi=fnd&pg=PR5&dq=cameron+metaphor+in+educational+discourse&ots=nGFH9g-DwQ&sig=Wy_haQy2OBA8bHpKv0mxsrLAll8

Clark, S. (2015). Vector space models of lexical meaning. *Handbook of Contemporary Semantics*, *10*, 9781118882139.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, *33*(4), 497–505.

Cressie, N., & Read, T. R. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 440–464.

Dodge, E., Hong, J., & Stickles, E. (2015). MetaNet: Deep semantic automatic metaphor analysis. *NAACL HLT 2015*, 40.

Droganova, K., & Medyankin, N. (2016). NLP pipeline for Russian: an easy-to-use web application for morphological and syntactic annotation. *Proceedings of the Annual International Conference "Dialogue"*. Presented at the Annual International Conference "Dialogue", Moscow.

Dunn, J. (2013a). Evaluating the premises and results of four metaphor identification systems. *International Conference on Intelligent Text Processing and Computational Linguistics*, 471–486. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-37247-6_38

Dunn, J. (2013b). What metaphor identification systems can tell us about metaphor-in-language. *Proceedings of the First Workshop on Metaphor in NLP*, 1–10. Retrieved from https://www.aclweb.org/anthology/W/W13/W13-09.pdf#page=11

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*(1), 1–24.

Ellis, N. C., & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, *7*(1), 188–221.

Evert, S. (2005). *The statistics of word cooccurrences: word pairs and collocations*.

Fauconnier, G., & Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.

Fenogenova, A., Kayutenko, D., & Dereza, O. (2015). *Mystem+*. Retrieved from http://web-corpora.net/wsgi/mystemplus.wsgi/mystemplus/

Fillmore, C. J. (1985). Syntactic intrusions and the notion of grammatical construction. *Annual Meeting of the Berkeley Linguistics Society, 11*, 73–86.

Fillmore, C. J. (1988). The mechanisms of "construction grammar". *Annual Meeting of the Berkeley Linguistics Society, 14*, 35–55.

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 501–538.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*.

Gandy, L., Allan, N., Atallah, M., Frieder, O., Howard, N., Kanareykin, S., … Argamon, S. (2013). Automatic Identification of Conceptual Metaphors With Limited Knowledge. *AAAI*. Retrieved from https://cps-xena.cps.cmich.edu/lgandy/automatic_metaphor.pdf

Gedigian, M., Bryant, J., Narayanan, S., & Ciric, B. (2006). Catching metaphors. *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, 41–48. Association for Computational Linguistics.

Goatly, A. (1997). *The language of metaphors* (Vol. 37). Retrieved from http://www.questia.com/library/communication/language-and-linguistics/grammar-and-word-use/metaphor

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Grady, J. (1997). *Foundations of meaning: Primary metaphors and primary scenes*.

Gurin, G., & Belikova, A. (2012). Metodika ocenki konvencional`nosti metaforicheskix vy`razhenij: ot intuitivistskix kriteriev k operacional`ny`m [A procedure for evaluating degree of conventionality of metaphor expressions: from intuition to operational criteria]. *Proceedings of the Annual International Conference “Dialogue”*, *1*, 187–197. Moscow, Russia.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in english* (1st ed.). Longman.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2–3), 146–162.

Heintz, I., Gabbard, R., Srivastava, M., Barner, D., Black, D., Friedman, M., & Weischedel, R. (2013). Automatic extraction of linguistic metaphors with lda topic modeling. *Proceedings of the First Workshop on Metaphor in NLP*, 58–66.

Hendricks, R. K., & Boroditsky, L. (2016). Emotional implications of metaphor: consequences of metaphor framing for mindset about hardship. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 1164–1169.

Herbelot, A., & Kochmar, E. (2016). ‘Calling on the classical phone’: a distributional model of adjective-noun errors in learners’ English. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 976–986.

Hovy, D., Srivastava, S., Jauhar, S. K., Sachan, M., Goyal, K., Li, H., … Hovy, E. (2013). Identifying metaphorical word use with tree kernels. *Proceedings of the First Workshop on Metaphor in NLP*, 52–57. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.362.39&rep=rep1&type=pdf#page=62

Johnson, C. (1996). Learnability in the acquisition of multiple senses: SOURCE reconsidered. *Annual Meeting of the Berkeley Linguistics Society*, *22*, 469–480.

Johnson, C. (1999). Metaphor vs. Conflation in the Acquisition of Polysemy: The Case of SEE.''. *Cultural, Psychological and Typological Issues in Cognitive Linguistics: Selected Papers of the Bi-Annual ICLA Meeting in Albuquerque, July 1995, 152*, 155. John Benjamins Publishing.

Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. *Language*, *39*(2), 170–210.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., … Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7–36.

Klebanov, B. B., & Flor, M. (2013). Argumentation-relevant metaphors in test-taker essays. *Proceedings of the First Workshop on Metaphor in NLP*, 11–20.

Klebanov, B. B., Leong, B., Heilman, M., & Flor, M. (2014). Different texts, same metaphors: Unigrams and beyond. *Proceedings of the Second Workshop on Metaphor in NLP*, 11–17.

Klebanov, B. B., Leong, C. W., Gutierrez, E. D., Shutova, E., & Flor, M. (2016). Semantic classifications for detection of verb metaphors. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, *2*, 101–106.

Klebanov, B. B., Shutova, E., & Lichtenstein, P. (2014). Proceedings of the Second Workshop on Metaphor in NLP. *Proceedings of the Second Workshop on Metaphor in NLP*. Presented at the Baltimore, MD. Retrieved from http://aclweb.org/anthology/W14-2300

Klebanov, B. B., Shutova, E., & Lichtenstein, P. (2016). Proceedings of the Fourth Workshop on Metaphor in NLP. *Proceedings of the Fourth Workshop on Metaphor in NLP*. Presented at the San Diego, California. Retrieved from http://aclweb.org/anthology/W16-1100

Kovecses, Z. (2010). *Metaphor: A Practical Introduction, 2nd Edition* (2e edition). Oxford; New York: Oxford University Press.

Kulagin, D. (2017). *Kartaslov*. Retrieved from https://github.com/dkulagin/kartaslov (Original work published 2017)

Kulagin, D. (2018). Otkry`taya semantika russkogo yazy`ka [The open semantics of the Russian Language]. Retrieved 5 March 2019, from https://habr.com/ru/post/434154/

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Kutuzov, A., & Kuzmenko, E. (2016). WebVectors: a toolkit for building web interfaces for vector semantic models. *International Conference on Analysis of Images, Social Networks and Texts*, 155–161. Springer.

Lakoff, G. (2008). *Women, fire, and dangerous things*. University of Chicago press.

Lakoff, G., Espenson, J., & Schwartz, A. (1991). *Master Metaphor List*. University of California at Berkely.

Lakoff, G., & Johnson, M. (1980a). *Metaphors We Live By* (2nd ed.). Chicago-London: The University of Chicago Press.

Lakoff, G., & Johnson, M. (1980b). The metaphorical structure of the human conceptual system. *Cognitive Science*, *4*(2), 195–208.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh* (Vol. 4). New york: Basic books.

Lambrecht, K. (1994). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents* (Vol. 71). Cambridge university press.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151–171.

Leong, C. W. B., Klebanov, B. B., & Shutova, E. (2018). A report on the 2018 VUA metaphor detection shared task. *Proceedings of the Workshop on Figurative Language Processing*, 56–66.

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins Publishing Company.

Lopukhina, A., & Lopukhin. (2017). Word Sense Frequency Estimation for Russian: Verbs, Adjectives, and Different Dictionaries. *Electronic Lexicography in the 21st Century. Proceedings of ELex 2017 Conference*, 267–280. Brno.

Lopukhina, A., Lopukhin, K., & Nosyrev, G. (2018). Automated Word Sense Frequency Estimation for Russian Nouns. In *Quantitative approaches to the Russian language* (pp. 79–94). Routledge.

Lyashevskaya, O., & Sharoff, S. (2009). *Chastotny`j slovar` sovremennogo russkogo yazy`ka na materialax Nacional`nogo korpusa russkogo yazy`ka / A frequency dictionary of modern Russian language on the basis of the Russian National Corpus*. Azbukovnik.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Mehlig, H. R. (1985). Semantika predlozheniya i semantika vida v russkom yazy`ke / The semantics of the sentence and the semantics of the aspect in the Russian language. In T. V. Bulygina & A. E. Kibrik (Eds.), *Novoe v zarubezhnoj lingvistike / Advances in international linguistics* (pp. 227–249). Moscow: Progress.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781*.

Mohler, M., Brunson, M., Rink, B., & Tomlinson, M. T. (2016). Introducing the LCC Metaphor Datasets. *LREC*.

Mohler, M., Rink, B., Bracewell, D. B., & Tomlinson, M. T. (2014). A Novel Distributional Approach to Multilingual Conceptual Metaphor Recognition. *COLING*, 1752–1763. Retrieved from http://www.aclweb.org/anthology/C14-1165

Narayanan, S. (1997). Embodiment in language understanding: Sensory-motor representations for metaphoric reasoning about event descriptions. *University of California, Berkeley: Unpublished Doctoral Dissertation*.

Neuman, Y., Assaf, D., Cohen, Y., Last, M., Argamon, S., Howard, N., & Frieder, O. (2013). Metaphor Identification in Large Texts Corpora. *PLOS ONE*, *8*(4), e62343. https://doi.org/10.1371/journal.pone.0062343

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100–108. Association for Computational Linguistics.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., … Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, *13*(2), 95–135.

Ovchinnikova, E., Israel, R., Wertheim, S., Zaytsev, V., Montazeri, N., & Hobbs, J. (2014). Abductive inference for interpretation of metaphors. *Proceedings of the Second Workshop on Metaphor in NLP*, 33–41. Retrieved from https://pdfs.semanticscholar.org/64c2/9feb317f54f38a4b61aa1a4b619cd3b90018.pdf#page=43

Panicheva, P., & Badryzlova, Y. (2017a). Distributional semantic features in Russian verbal metaphor identification. *Computational Linguistics and Intellectual Technologies*, *1*, 179–190.

Panicheva, P., & Badryzlova, Y. (2017b). Distributional semantic features in Russian verbal metaphor identification. *Computational Linguistics and Intellectual Technologies*, *1*, 179–190. Moscow.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Dubourg, V. (2011). *Scikit-learn: Machine Learning in Python Journal of Machine Learning Research*.

Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., McRae, K., & Spivey, M. J. (2013). Computational Grounded Cognition: a new alliance between grounded cognition and computational modeling. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00612

Pragglejaz Group. (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, *22*(1), 1–39.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Rosen, Z. (2018). Computationally Constructed Concepts: A Machine Learning Approach to Metaphor Interpretation Using Usage-Based Construction Grammatical Cues. *Proceedings of the Workshop on Figurative Language Processing*, 102–109.

Russian National Corpus. (n.d.). Retrieved 9 May 2019, from http://www.ruscorpora.ru/en/index.html

RusVectōrēs: semanticheskie modeli dlya russkogo yazy`ka [RusVectōrēs: semantic models for the Russian language]. (n.d.). Retrieved 25 April 2019, from RusVectores website: https://rusvectores.org/ru/

Schmid, H. (1994). Probabilistic Part-of-speech Tagging Using Decision Trees. *International Conference on New Methods in Language Processing*. Presented at the International Conference on New Methods in Language Processing, Manchester, UK. Retrieved from http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf

Segalovich, I. (2003). A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. *MLMTA*, 273–280. Citeseer.

Sense frequecies with Russian Active Dictionary. (n.d.). Retrieved 9 May 2019, from http://sensefreq.ruslang.ru/

Shutova, E., Kiela, D., & Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features. *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 160–170. Retrieved from https://pdfs.semanticscholar.org/1506/645e42043637607b3acd0c42b562d5a336b1.pdf

Shutova, E., Klebanov, B. B., & Lichtenstein, P. (Eds.). (2015). *Proceedings of the Third Workshop on Metaphor in NLP*. Retrieved from http://www.aclweb.org/anthology/W15-14

Shutova, E., Klebanov, B. B., Tetreault, J., & Kozareva, Z. (2013). Proceedings of the First Workshop on Metaphor in NLP. *Proceedings of the First Workshop on Metaphor in NLP*. Presented at the Atlanta, Georgia. Retrieved from http://aclweb.org/anthology/W13-0900

Shutova, E., Sun, L., Gutiérrez, E. D., Lichtenstein, P., & Narayanan, S. (2017). Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning. *Computational Linguistics*, *43*(1), 71–123.

Shutova, E., & Teufel, S. (2010). Metaphor Corpus Annotated for Source-Target Domain Mappings. *LREC*, *2*, 2–2. Retrieved from http://lexitron.nectec.or.th/public/LREC-2010_Malta/pdf/612_Paper.pdf

Siegal, S. (1956). *Nonparametric statistics for the behavioral sciences*. McGraw-hill.

Steen, G., Herrmann, B., Kaal, A., Krennmayr, T., & Pasma, T. (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Amsterdam; Philadelphia, PA: John Benjamins Publishing Company.

Steen, G. J., A.G.Dorst, J.B.Herrmann, A.A.Kaal, T.Krennmayr, & T.Pasma. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*. Amsterdam: John Benjamins.

Strzalkowski, T., Broadwell, G. A., Taylor, S., Feldman, L., Yamrom, B., Shaikh, S., … others. (2013). Robust extraction of metaphors from novel data. *Proceedings of the First Workshop on Metaphor in NLP*, 67–76. Retrieved from http://www.cl.cam.ac.uk/~es407/papers/meta4NLP2013.pdf#page=77

Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors We Think With: The Role of Metaphor in Reasoning. *PLoS ONE*, *6*(2), e16782. https://doi.org/10.1371/journal.pone.0016782

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., & Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, *1*, 248–258.

Turney, P. D., Neuman, Y., Assaf, D., & Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 680–690. Retrieved from http://dl.acm.org/citation.cfm?id=2145511

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Varela, F. J., Thompson, E., & Rosch, E. (2017). *The embodied mind: Cognitive science and human experience*. MIT press.

Veale, T. (2018). The “default” in our stars: Signposting non-defaultness in ironic discourse. *Metaphor and Symbol*, *33*(3), 175–184. https://doi.org/10.1080/10926488.2018.1481262

Veale, T., Shutova, E., & Klebanov, B. B. (2016). Metaphor: A Computational Perspective. *Synthesis Lectures on Human Language Technologies*, *9*(1), 1–160. https://doi.org/10.2200/S00694ED1V01Y201601HLT031

Vendler, Z. (1957). Verbs and times. *The Philosophical Review*, *66*(2), 143–160.

Wilks, Y. (1978). Making preferences more active. *Artificial Intelligence*, *11*(3), 197–223.

Wu, C., Wu, F., Chen, Y., Wu, S., Yuan, Z., & Huang, Y. (2018). Neural Metaphor Detecting with CNN-LSTM Model. *Proceedings of the Workshop on Figurative Language Processing*, 110–114.

Yevgenyeva, A. (Ed.). (1981). *Dictionary of the Russian Language* (2nd ed., Vols 1–4). Moscow: Academy of Sciences of the USSR; Russian Language Institute.