ISSN 2007-9737

Distributional Word Vectors as Semantic Maps Framework

Amir Bakarov

National Research University Higher School of Economics, Moscow, Russia

amirbakarov@gmail.com

Abstract. Distributional Semantics Models are one of the most ubiquitous tools in Natural Language Processing. However, it is still unclear how to optimize such models for specific tasks and how to evaluate them in a general setting (having ability to be successfully applied to any language task in mind). We argue that benefits of intrinsic distributional semantic models evaluation could be questioned since the notion of their "general quality" possibly does not exist; distributional semantic models, however, can be considered as a part of Semantic Maps framework which formalizes the notion of linguistic representativeness on the lexical level.

Keywords. Word embeddings, distributional word vectors, semantic maps.

1 Introduction

The Semantic Maps framework in linguistics aims to describe patterns of multifunctionality of grammatical units without grounding to monosemic and polysemic analyses [38]. The core concept of this framework is a semantic map, geometrical representation of grammatical functions (such as uses, meanings, and contexts of grammatical morphemes) as interlinked constituents a so-called "semantic space", a structure that implies graph theory mechanisms and claims to generalize the configuration of functions shown by the map across linguistic phenomena and different languages.

This structure could be viewed as a representation of conceptual similarity between different semantic functions [39], certain scholarly studies, though, do not impose such attribution, and interpret Semantic Maps as a compact description of attested variation, imposing a question of whether this framework may reflect extra-cognitive factors (diachronic or communicative) [37].

Semantic Maps are constructed with a core principle of "contiguity / connectivity requirement" in mind, functions that are often associated with one and the same linguistic expression are represented as nodes adjacent to each other, or as a contiguous region in a semantic map [73], but it does not imply that one and the same linguistic expression represented through an association with several nodes should be analyzed as polysemic.

Therefore, Semantic Maps claim that multifunctionality of a gram occurs only when the various functions of the gram are similar, for example, as one of the possible application of Semantic Maps is separation of polysemy from accidental homonymy, where formally identical elements have unrelated meanings [28].

Ideally, a complete theory of grammatical meaning would allow us to deductively leverage Semantic Maps for deriving language-independent functions as well as their relative positions at the map structure, as functions in Semantic Maps are distributed in a way that allows each gram from each language to occupy a contiguous area on a map.

However, given the data of only one language, we can not be sure which functions to represent on the map in the first place [88]. All in all, Semantic Maps have become a popular method in grammatical typology, being used for capturing both synchronic facts and patterns of development [41].

One of the types of Semantic Maps on which we focus in this article is called Probabilistic Semantic Maps, a way of constructing Semantic Maps through statistical methods based on correspondence analysis of relative occurrences of particular linguistic expressions in different contexts across one or multiple languages [88]. “Semantic space” operated by this type of Semantic Maps is expressed topologically by closeness of nodes in the Semantic Map graph. WordNet [48] is one of the most well-known examples of Probabistic Semantic Maps.

It is manually constructed using heuristic judgments on the similarity concepts as a medium of multifunctionality. Recent scholarly studies propose an alternative to the manual construction of such maps with a Distributional Semantics theory, a context-based, non-compositional approach to meaning [40], following the claim that the meaning of a word can be determined based on patterns of co-occurrence in a corpus [53].

The fundamental assumption in Distributional Semantics is that the word meaning is distributed across contexts of its use, and lexical representations are quantitative functions of their global distributions, which can be viewed as so called Word Vectors.

Given metric as a measure of similarity of words corresponding to given vectors, one can use it as a proxy for semantic relations between corresponding words. This metric can be formally represented with an any kind of similarity measure between vectors, like cosine similarity:

$$\cos(x,\ y) = \frac{x \cdot y}{||x|| \cdot ||y||}. \tag{1}$$

Here x and y are compared vectors. Distributional Semantics has a number of features that make it different from other semantic theories and made it able to become the most ubiquitous semantic concept nowadays:

1. **Distributional word representations** are context-sensitive: linguistic contexts in which words occur construct their semantic constitution.
2. **Distributional word representations** are inherently distributed, i.e., captured word meaning lies in overall distributional history of this word rather than in certain set of explicitly observable features.
3. **Distributional word representations** are gradual, so the captured meaning differs not merely for the contexts they appear, but also for how saliently these contexts describe the combinatorial behavior of these words.

Far from being a “theory-neutral” approach to semantics, Distributional Semantics has been used to test linguistic hypotheses, and there is some evidence supporting the view that semantic associations and textual co-occurrences are related [96]. Distributional Semantics is particularly well suited to describing those aspects of meaning that interact with syntax, such as argument structure.

Usually scholars distinguish two taxonomic classes of distributional semantic models [13]. The first one is based on explicit counts of word co-occurrences in a corpus. Such counts can be done by finding all word-per-word occurrences and measuring the degree of mutual information inheriting from this connection.

Ubiquitous example of a distributional semantic model based on such counts is a model of *Latent Semantic Analysis* [92] which actually builds sparse word-word matrix for pointwise mutual information of word co-occurrences and applies dimensionality reduction on this matrix. This class of distributional semantic models is therefore called count-based distributional semantics.

Models from the second taxonomic class are based on sampling the training corpus with a sliding window, so each word is initialized with a feature vector which values are optimized to accurately predict sequence of words in the corpus given an input sequence of words (language modeling). Since usually this task is resolved with the help of artificial neural networks, such distributional semantic models are usually called neural-based (or prediction-based) distributional semantic models.

Meaning representations captured as input weights of these networks are called word

embeddings [35]. Despite such taxonomy is nominal, several recent works have proven the effectiveness of predictive models against count-based models [13].

Nowadays the neural-based architectures gained the most popularity in the community along with such algorithms as *Continuous Skip-Gram* or *Continuous Bag-of-Words* [106], *Global Vectors* [118], *FastText* [24], and others. Despite distributional hypothesis gained most attention after Harris's work [71], the pre-requisites and early versions of this hypothesis were formulated by other post-Bloomfieldian American structuralists: Martin Koos, Charles Hockett and George Trager [125].

To this end, distributional hypothesis has much connection with structuralist hypothesis, or formal approach to linguistic in general [50]. As in structural models, words in distributional hypothesis are defined according to their features in the lexicon, and the meanings are defined by contrasts in these sets of features: in distributional hypothesis words' contexts of use play the role of these features, while structuralist hypothesis relies on manually handcrafted properties.

This leads to lower linguistic motivation of distributional hypothesis comparing to the structuralist one: unsupervised feature construction is more convenient for downstream tasks, but they can lack an intrinsic meaning [25]. This issue of intrinsic vagueness grows from the application purpose of distributional hypothesis: as [125] puts it, the structuralist distributional procedure was originally introduced for phonemic analysis, and only after few time turned into a general methodology able to be applied to every linguistic level.

This procedure was a way for linguists to ground their analyses on firm methodological base, avoiding any argument based on meaning as an identity criterion for linguistic elements [64]. But mapping distributional hypothesis to word studies can go against semantic theories relying on precise identity criteria for semantic content of words since distributional semantics build this criteria on a generalization of paradigmatic relations built upon on linguistic distributions.

For example, we can formulate distributional hypothesis as the ability of the degree of semantic similarity between two linguistic expressions A and B to be considered as a function of the similarity of linguistic contexts in which A and B can appear [96]. One can observe that this actually inherits Bloomfield's refusal of meaning as an linguistic explanans [23], defining meaning as a similarity of words' distribution (in other words, it supports the idea that an exploration of a number of contexts of a word can evidence some of its semantic properties). But while Bloomfield assumes that research of meaning goes beyond linguistic research, the distributional hypothesis implicitly puts a solid empirical (statistical) foundation for meaning analysis.

However, we assume that meaning becomes a part of empirical studies only at those aspects that can be defined through distributional analysis procedures. From the position of cognitive linguistics, many more aspects of meaning are not ever fixed in written traditions [63]. Through the view of conceptual hypothesis and prototypical hypothesis (or functional approach in general), the distributional theory also could be motivated through the same methods that appear as ingredients of human conceptualization (particularly, the linguistic contexts).

According to [93], contexts intrinsically embody conceptual representation of aspects of the world. Such representations commonly propose a functional explanation in terms of the principles governing the process of conceptualizing the word. Therefore linguistic distributions and meaning proposed by them are explained in many cases by embodied conceptualisation.

Certain psycholinguists, e.g. [109], assume that repeated encounters of a word in various linguistic contexts eventually determine the formation of a contextual representation. This view on meaning highly relates with that Wittgenstein puts as "meaning of word is its use in the language" [143].

However, formal (Harris') and functional (Miller's) views on distributional theory have a few in common. While Harris puts distributional analysis as a purely empirical method of linguistic research, Miller assumes a cognitive background of meaning

and claims that it goes beyond purely statistical investigation. To this end, linguistic motivation of distributional semantic models relies on dichotomy of these theories, and can be explained from both these sides (either as be criticized).

If we want to support the claim of Distributional Semantics as a legit linguistic framework that can be viewed from the perspective of Semantic Maps theory, we should accept assumptions of one of these views. From the structural views, we must posit that language is a system of inter-related units and structures and that every unit of language is related to the others within the same system. From the conceptual views, we should assume that mental representations which encode the human understanding of the world contain the primitive conceptual elements of which meanings are built.

However, despite all this recent progress, structural and conceptual theories are still much more heuristics than practice and yet lack a strong experimental basis. Therefore, the possibility to support the legitimacy of Distributional Semantics as a linguistic framework can be doubted.

To this end, we suggest to turn the investigations of linguistic legitimacy of distributional semantics to the computational side, which proposes more objective framework. As probabilistic semantic maps can be viewed as modeling the semantics of linguistic diversity (and they do so to the extent that the sample, which is an underlying typological database, is representative of the populatio, which is the entire linguistic diversity), we can pose a general question is whether semantic maps based on linguistic data can model universal semantic space, claimed to be the ultimate aim of Semantic Maps framework in the beginning of this article.

If semantic space is both mental and universal, it must be both comprehensive and robust. Robust means that different samples of languages and of semantic functions are assumed to yield highly similar maps representing the full range of semantic diversity encountered in natural languages. Comprehensive means that all semantic categories encountered in the database must be well-represented [141].

While being more focused on “the semantics of individual lexical items, their configurations in lexical field or individual processes of word formation” rather than on “typologically relevant features in the grammatical structure of the lexicon” [95], we assume that the primary benchmark for distributional semantic models as semantic maps can be proposed from the perspective of evaluation of their features as models of lexicon. Existing approaches to evaluation of distributional semantic models divide on perspectives of extrinsic evaluation (evaluation on downstream tasks) and intrinsic evaluation (evaluation of inner properties of the models) [128].

Former methods are based on the ability of distributional word vectors to be used as the feature vectors of supervised machine learning algorithms used in one of various downstream natural language processing tasks. On the opposite, methods of intrinsic evaluation are experiments in which word vectors are compared with human judgements on words relations. Manually created sets of words are often used to get human assessments, and then these assessments are compared with word vectors.

Intrinsic evaluation relies on in vivo experiments to obtain human judgments from assessors. Such an estimate could be used as an absolute measure of the quality of word vectors since it reports the similarity of lexical semantics inferred by a distributional semantic model to the lexical semantics determined by humans. To this end, I consider distributional semantic model representative of language only in case they demonstrate decent evaluation performance.

Intrinsic evaluation approaches like the “word semantic similarity task” (will be covered in more detail in the next section) do not use such formalizable a strict notion of the model’s performance. From an intrinsic evaluation perspective, word embeddings are usually assessed using our (humans’) understandings of relationships between words (and other lexical units), for example, by collecting human annotations of a so-called “word semantic similarity”.

Usually, intrinsic evaluation relies on psycholinguistic tasks which collect human judgments on the “gold standard” of different properties of the lexicon. These tasks (experiments) are conducted either in the laboratory

with a limited set of examinees (**judgments collected in-house**) or on crowd-sourcing Web platforms like Mechanical Turk, attracting an unlimited number of participants (**judgments collected through crowd-sourcing**) [98].

Sometimes the assessors are asked to evaluate the quality of word embeddings directly, for instance, when different models produce different judgments on word relations, and the task of an assessor is to tell which model works better. This type of intrinsic evaluation is called **comparative intrinsic evaluation** [128], in opposite to the regular or "absolute" intrinsic evaluation, it allows not to estimate the absolute quality of the word vectors, but to find the most adequate vectors in a given set.

As it was mentioned, unlike "extrinsic" evaluation, the "intrinsic" approach tries to assess a more general notion of word embeddings performance (particularly, can DSMs be used as a proper formalization of a lexicon?) but at the same time, it relies on less formalizable and more vague concepts. Particularly, is unclear which type of relationships the word embeddings should reflect (synonymy? co-hyponymy? something else), and if the model takes into account not the type of relationships that we know (like synonymy), but other types of relationships, how should we assess it [47].

Scholar works on this topic tend to face various methodological problems, such as lack of proper test sets (resulting in adjusting the models to the data trying to increase their quality) or absence of the statistical significance tests. One of the main issues with most of the scholarly research on this topic is that there is no strict definition of an evaluation method in the field of distributional semantics (after all, if the notion of word meaning could not be even defined properly, how the notion of its modeling evaluation could be defined?).

Therefore, we consider by the method of word embeddings evaluation any way or attempt of finding a link (correlation) between a DSM and **any** data that hypothetically could carry information about lexical semantics. The evaluation representativeness obviously depends on the degree of plausibility of the hypothetical amount of lexicon information in the data one tries to use for evaluation, but the general intuition is that we are not able to strictly evaluate this amount.

2 Empirical Benchmarks

2.1 Semantic Similarity

The most well-known benchmark of **word semantic similarity** directly assesses the ability of DSMs to report representative distance between the word vectors in terms of the ability of this distance to be grounded to human assumptions on that distance between corresponding words.

For example, if the so-called "distance" between *cup* and *mug* (defined in a continuous interval $0, 1$) predicted by the model is 0.8, then we assume that the distance is reported correctly by the model of the human assessor asked to estimate the "distance" between these words (whatever it means depending on the annotation guidelines) outputs a similar value.

These distances, both DSM's and human's one, are collected on a range of pairs of words, and we expect to find a meaningful correlation between these two sets (usually, more than one assessor is used for the sake of reliability of the provided scores). Having two different models, we consider the better model the more correlated are the predictions [13].

Word similarity benchmark is also one of the oldest ones, its roots go back to 1965 when the first experiments on human judgments on word semantic similarity were conducted to test the distributional hypothesis from the psychology perspective [124] (in 1978 a similar work was carried out in [112]).

Despite the strong psycholinguistic background of this method, it is one of the most frequently criticized in the community, obviously for subjectivity and vagueness [54, 47, 18], there are a lot of potential linguistic, psychological [85] and social factors [117], which could introduce bias in the assessments [61].

The task is also very much dependent on possible connotations in the word lists [98], and the ambiguity of the overall assessment, different works propose different definitions of semantic similarity, while some scholars define

it as co-hyponymy (like the relation between the words *machine* and *bicycle*) [137], others define it as synonymy (like in a word pair *mug* and *cup*) [77].

It was also argued that the notion of semantic similarity inherits not only semantic connections of words but also some morphological and graphemic features of word representations [87].

Among other criticized features of word semantic similarity, there is also the lack of correlation between these human assessments and the performance of word embeddings on extrinsic methods [33, 132], the low inter-rater agreement between annotators [77], the factor of assessors getting tired when annotating a large number of pairs [27], poor ability of numerical labels to fully describe all types of relations between words (it is suggested that it will be better to describe the degree of word similarity in a natural language [108]), and the misconduct of thematic roles relations [44].

It is also unclear whether such embeddings reflect enduring properties of language or if they are sensitive to inconsequential variations in the source documents [6, 8]. Datasets:

1. **SimVerb-3500**, 3 500 pairs of verbs assessed by semantic similarity (that means that pairs that are related but not similar have a fairly low rating) with a scale from 0 to 4 [59].
2. **MEN** (acronym for Marco, Elia and Nam), 3 000 pairs assessed by semantic relatedness with a discrete scale from 0 to 50 [27].
3. **RW** (acronym for Rare Word), 2 034 pairs of words with low occurrences (rare words) assessed by semantic similarity with a scale from 0 to 10 [100].
4. **SimLex-999**, 999 pairs assessed with a strong respect to semantic similarity with a scale from 0 to 10 [77].
5. **SemEval-2017**, 500 pairs assessed by semantic similarity with a scale from 0 to 4 prepared for the *SemEval-2017 Task 2* (*Multilingual and Cross-lingual Semantic Word Similarity*) [30]. Notably, dataset contains not only words, but also collocations (*e.g. climate change*).
6. **MTurk-771** (acronym for Mechanical Turk), 771 pairs assessed by semantic relatedness with a scale from 0 to 5 [69].
7. **WordSim-353**, 353 pairs assessed by semantic similarity (however, some researchers find the instructions for assessors ambiguous with respect to similarity and association) with a scale from 0 to 10 [51].
8. **MTurk-287**, 287 pairs assessed by semantic relatedness with a scale from 0 to 5 [122].
9. **WordSim-353-REL**, 252 pairs, a subset of WordSim-353 containing no pairs of similar concepts [3].
10. **WordSim-353-SIM**, 203 pairs, a subset of WordSim-353 containing similar or unassociated (to mark all pairs that receive a low rating as unassociated) pairs [3].
11. **Verb-143**, 143 pairs of verbs assessed by semantic similarity with a scale from 0 to 4 [12].
12. **YP-130** (acronym for Yang and Powers), 130 pairs of verbs assessed by semantic similarity with a scale from 0 to 4 [144].
13. **RG-65** (acronym for Rubenstein and Goodenough), 65 pairs assessed by semantic similarity with a scale from 0 to 4 [124].
14. **MC-30** (acronym for Miller and Charles), 30 pairs, a subset of RG-65 which contains 10 pairs with high similarity, 10 with middle similarity and 10 with low similarity [109]. Also, there is a subset of MC-30 called **MC-28** which excludes 2 pairs not represented in WordNet [123].

2.1.1 Synonym Detection

The so-called Synonym Detection task is very close to the previously described task of Semantic Similarity, but while it also assesses the ability of DSMs to provide reliable distances between words, it does not rely on an absolute degree of similarity in terms of a scalar value. Instead, we assume that

we can do the thing by finding the most similar word relative to a set of other words.

So, given a word a and a set $K = b_1, b_2, b_3$, the task is to find b_i which is the most synonymous (semantically similar in terms of the word semantic similarity task) to a [13].

For example, for the target *levied* one must choose between *imposed* (correct), *believed*, *requested* and *correlated*. The task of a DSM is to find the word vector with the smallest distance to the vector of the specified word.

Taking into account all the criticism of the word semantic similarity method, moving from the absolute measure to the relative measure could probably exclude a lot of problems of this task (score bias, lack of assessments interpretability, etc.).

On the other hand, the creation of a dataset for evaluation in this task is more complicated and raises certain new questions (for example, how to properly choose the words to form the set K). Datasets that could be used for evaluation on this task presented in a form of 5-word tuples in which one word is a target word, and 4 words are potential synonyms where the only one is a correct answer:

1. **RDWP** (acronym for Reader's Digest Word Power Game; also mentioned as RD-300), 300 synonym questions (5-word tuples) [82].
2. **TOEFL** (acronym for Test of English as a Foreign Language), 80 questions [92].
3. **ESL** (acronym for English as a Second Language), 50 questions [134].

2.2 Word Analogy

The task of Word Analogy (in some works being also called *analogical reasoning*, *linguistic regularities* and *word semantic coherence*) implies the intuition that the arithmetic operations in a word vector space should have a common sense reasoning.

Given a set of three words, a, $a*$ and b, the task is to identify such word $b*$ that the relation b:$b*$ is the same as the relation a:$a*$ [133, 119, 13]. For instance, having $a = Paris$, $b = France$, $c = Moscow$, the target word would be $Russia$ since the relation $a : b$ is $capital : country$, so one needs need to find the capital of which country is *Moscow*. There are different metrics that can be used in this benchmark, though:

— *3CosAdd* (and a similar metric *3CosMul*) proposed in the original *Word2Vec* paper is based on arithmetic operations in vector space (addition and multiplication of cosine distances) [107].

— *PairDir* modifies *3CosAdd*, taking into account the direction of the resulting vectors in these operations [97].

— *Analogy Space Evaluation* metric compares the distances between words directly without finding the nearest neighbors [32].

This task was also criticized and investigated at [110]. A theoretical investigation of analogy phenomena of word vectors was presented in [60]. There was a concept of temporal word analogies also introduced [131].

[52] also gives much attention to the problem of analogy solving. We also provide a list of datasets which could be used for evaluation on this task. As [62] notes, datasets designed for *semantic relation extraction task* could also be used to compile a word analogy set.

In this case, it also worth looking at the *Lexical Relation* set which is a compilation of different semantic relation datasets including *BLESS* [16] (12 458 word pairs with a relation comprising 15 relation types) [140] and the *Semantic Neighbors* set (14 682 word pairs with a relation comprising 2 relation types, meaningful and random) [115].

1. **WordRep**, 118 292 623 analogy questions (4-word tuples) divided into 26 semantic classes, a superset of *Google Analogy* with additional data from WordNet [57].
2. **BATS** (acronym for Bigger Analogy Test Set), 99 200 questions divided into 4 classes (*inflectional morphology*, *derivational morphology*, *lexicographic semantics* and *encyclopedic semantics*) and 10 smaller subclasses. [62].

3. **Google Analogy** (also called Semantic-Syntactic Word Relationship Dataset), 19 544 questions divided into 2 classes (*morphological relations* and *semantic relations*) and 10 smaller subclasses (8 869 semantic questions and 10 675 morphological questions) [105].
4. **SemEval-2012**, 10 014 questions divided into 10 semantic classes and 79 subclasses prepared for the *SemEval-2017 Task 2* (*Measuring Degrees of Relational Similarity*) [86].
5. **MSR** (acronym for Microsoft Research Syntactic Analogies), 8 000 questions divided into 16 morphological classes [107].
6. **SAT** (acronym for Scholastic Aptitude Test), 5 610 questions divided into 374 semantic classes [136].
7. **JAIR** (acronym for Journal of Artificial Intelligence Research), 430 questions divided into 20 semantic classes. Notably, dataset contains not only words but collocations (like *solar system*) [135].
8. New analogical reasoning dataset [72].

2.3 Thematic Fit

The method of Thematic Fit (also called *selectional preference* in [13]) is to separate different thematic roles of arguments of a predicate and to find how well the word embeddings could find most semantically similar noun for a verb that is used in a specific role.

For humans, a certain verb could cause a person to expect that a certain role must be filled with a certain noun (e.g., for the argument *to cut* the most expected argument in the *object* role is *pie*) [127].

Experiments propose assessments of adequacy score of the tuple $\{$verb, noun, thematic role$\}$ (for example, *people eat* is more common phrase than *eat people*, so the pair *people* and *eat* would have the higher score) [139].

Some researchers consider another variation of this method, proposing the task of assessing a pair of words n (noun) and v (verb) by the most frequent role in which n used with v (e.g., pair $people, eat$ would be classified as the *subject* since it is more common to use $people$ as a subject with that verb) [15].

It is unclear, though, which method of obtaining an embedding for a thematic role to distinguish different roles of the argument is the most adequate, some researchers propose a method of vectorization of "slots" for certain thematic roles, which are obtained by averaging several most frequent nouns encountered in a given role [15].

1. **MSTNN** (abbreviation mentioned in [127]), 1 444 *verb-object-subject* pairs [103].
2. **GDS** (acronym for Greenberg, Sayeed and Danberg), 720 *verb-object* pairs. The dataset is additionally divided into a subsample containing only polysemous verbs (*GDS-poly*) and a subsample containing monosemous verbs (*GDS-mono*) [65].
3. **F-Inst & F-Loc** (acronym for Ferretti-Instrument and Ferretti-Location), 522 verbs pairs which are split to a subset of 248 verbs with associated *instruments* (*F-Inst*) and a subset of 274 verbs with associated *locations* (*F-Loc*) [49].
4. **P07** (acronym for Pado), 414 *verb-object-subject* pairs [114].
5. **UP** (acronym for Ulrike and Pado), 211 *verb-noun* pairs, the set of roles is unlimited [113].
6. **MT98** (acronym for McRae and Tanenhaus), a subset of 200 verbs from *MSTNN* where each verb has two nouns, one is a good subject, but a bad object, and one which is a good object, but a bad subject [104].

2.4 Concept Categorization

The method of Concept Categorization assesses a word vector space's ability to be split into distinguishable categories, i.e., clusters. Given a set of words, we want to map each word into a meaningful category which can have common sense reasoning (for example, for words *dog*,

elephant, *robin*, *crow*, the first two make one cluster which is *mammals* and the last two form another second cluster which is *birds*; the cluster name is not necessary to be formulated) [13].

Lexical-typological research typically asks questions such as how languages categorize particular domains (human body, kinship relations, color, motion, perception, etc.) by means of lexical items, what parameters underlie categorization, whether languages are completely free to "carve up" the domains at an infinite and arbitrary number of places or whether there are limits on this, and whether any categories are universal (e.g., say 'relative', 'body', or 'red').

The critique of such a method addresses the question of either choosing the most appropriate clustering algorithm or choosing the most adequate metric for evaluating clustering quality.

A different way to approach this evaluation method was introduced in the works related to categorical modularity, which is a graph modularity metric based on the k-nearest neighbor graph constructed with embedding vectors of words from a fixed set of semantic categories, in which the goal is to measure the proportion of words that have nearest neighbors within the same categories [31].

The underlying principle is that in good embeddings, words in the same semantic category should be closer to each other than to words in different categories.

The authors quantify this by building the k-nearest neighbor graph with a fixed set of words' semantic categories and computing the graph's modularity for a given embedding space. Modularity measures the strength of division of a graph with densely connected groups of vertices, with sparser connections between groups [55]. The datasets for the Word Categorization task are presented with sets of words classified into a number of certain categories.

1. **BM** (acronym for Battig and Montague), 5 321 words divided into 56 categories [17].
2. **AP** (acronym for Almuhareb and Poesio), 402 words divided into 21 categories [4].
3. **BLESS** (acronym for Baroni and Lenci Evaluation of Semantic Spaces), 200 words divided into 27 semantics classes [16]. Despite the fact that BLESS was designed for another type for evaluation, it is also possible to use this dataset in a word categorization task, as in [83].
4. **ESSLLI-2008** (acronym for the European Summer School in Logic, Language and Information), 45 words divided into 9 semantic classes (or 5 in less detailed categorization); the dataset was used in a shared task on a *Lexical Semantics Workshop on ESSLI-2008* [14].

2.5 Outlier Word Detection

This method of Outlier Word Detection evaluates the same feature of word embeddings as the word categorization method (it also proposes clustering), but the task is not to divide a set of words into a certain amount of clusters, but to identify a semantically anomalous word in an already formed cluster (for example, for a set $\{$orange, banana, lemon, book, orange$\}$ which are mostly fruits, the word *book* is the outlier since it is not a fruit) [29].

Some researchers propose a very similar method called *evaluation of coherence in semantic space*. The idea of this method is, given a set of three words – word a, the two words a_1 and a_2 which are the closest to a in an embedding space are found, – a word b is chosen randomly from the model's dictionary (this word probably would not be so semantically similar to a), and the task of a human assessor is to correctly identify b (the outlier) in the set a, a_1, a_2, b [128]. The more words are identified correctly, the better is the model.

1. **8-8-8 Dataset**, 8 clusters, each is represented by a set of 8 words with 8 outliers [29].
2. **WordSim-500**, 500 clusters, each is represented by a set of 8 words with 5 to 7 outliers [20].

3 "Subconcsious" Experimental Evaluation Tasks

As we mentioned in the previous section, because the notion of DSMs quality is not bounded only by standard benchmark performance, extending to the territory of a more global question of building a lexicon model, we also attempt to overview experimental evaluation tasks that might not be industry applicable (yet), but which can provide important insights from linguistic and scientific points of view on distributional semantics. Moreover, with the recent trends in the community, these methods start to get out of the "experimental" zone and started to get more attention from different researchers both from cognitive sciences and from language technology [7].

Later in this section, we describe different ways of collecting cognitive data and their application to DSMs evaluation in more detail.

3.1 Semantic Priming

A semantic priming behavioral experiment is based on a hypothesis that a human should read a word faster if it is preceded by a semantically related word (which can introduce an association in a brain). Within the experiment, the time of reading a specified word a (called the *target word*) is compared with the time required to read it when it occurs after a word b_1 and with the time required to read it in a case it occurs after a word b_2.

If the reading time of the word b_1 is lower than the reading time of the word b_2, than the word b_1 is claimed to be semantically related to a (b_1 is called *prime*, or *prime word*, or *stimulus word*) [46, 9]. Reading time is tracked using eye-tracking or safe-paced reading [101, 94], [84, 76, 102, 126].

The most notable dataset used for semantic priming experiments is the *Semantic Priming Project*, containing 6 337 pairs of words. The data is collected from 768 subjects for 1 661 target words. Every word pair is presented in four versions: first, depending on the time interval on the demonstration of the target and non-target words which is 70 and 200ms (this interval is called *stimulus onset asynchronies, SOA*), and, second, depending on the task for the priming, naming task or lexical decision task [79].

3.2 Functional Magnetic Resonance Imaging

One of the most ubiquitous ways to analyze neural activity in a human brain is functional magnetic resonance imaging (functional MRI, fMRI), which records changes in blood level on the brain cortex (bloodoxygen-level-dependent (BOLD) responses), while the brain is presented with certain stimuli. BOLD responses are commonly represented as dense 4-D arrays of the measured data, where time series of the blood flow-related activity measured in tens of thousands of voxels (which are small areas of size equal to approximately 2x2x2 mm^3) are measured across the brain.

These excitations are hypothesized to be elicited primarily by the presented stimulus (with minor background contamination due to respiration, heartbeat, or movement). Hypothetically, the stimuli find their representations in these voxel patterns, and the works aiming to map DSMs data and fMRI data usually rely on matching these two types of data in different ways, particularly, to use word embeddings to try to predict voxel's activation [80, 78].

The evaluation method is based on using as a gold standard the data of fMRI experiments which measures changes associated with blood flow in certain parts of the brain by fixating regions of the blood flow at certain time intervals (once a second, for instance).

The idea is that the blood flow and the neuronal activation patterns correlate, so one could identify parts of the brain that are activated. In the field of neurolinguistics, reading or listening to the text is usually considered to be a stimulus for this activity. The obtained data is presented as a set of voxels reporting the level of neuronal activity in different small parts of the brain.

It is not clear how to obtain data on reading single words, since the minimum time interval on fixating blood flow is about 1 second; some researchers try to train a regression model to compute the average brain activation vectors for each word or to use aggregate statistics to obtain

vector representations of fMRI data using it is as a gold standard [130, 1].

One could try to use *StudyForrest* [70] dataset which offers data on listening to the audio track of the "Forrest Gump" movie in German, or the *Word Processing* dataset which contains readings for various natural language words on English [43]. Most of the studies, though, do not try to compare different word embeddings models to each other, but just try to figure out whether they are capable to encode abstract information [138].

3.3 Electroencephalography

Electroencephalography (EEG) records the electrical activity of the brain, and the idea is that the amplitude of the impulses in the brain that occur on words (such response is called N400, it is an early response elicited by every word of a sentence) stores information about lexical semantics since the interpretation of the response is usually generalized by the hypothesis that the worse the word fits the context (which could be both sentence context and word context), the higher is the amplitude of the signal.

The amplitude differences of a tuple of words are able to be simulated through the average cosine distances of word embeddings, so it is hypothetically could be used as a gold standard data for evaluation [116, 45, 129].

3.4 Eye Movement Data

This evaluation method is based on using as a gold standard the data of human eye movement obtained. Such data could be obtained through an instrument called *eye-tracker* which tracks the movement of a pupil and a time of fixation on certain words while a person reads text from the computer screen, and such data hypothetically could carry some information about lexical semantics.

The eye-tracker assigns to each word a set of features reporting characteristics of its reading: how many milliseconds the gaze was fixated on this word, how many times the gaze returned to it, etc. Such feature sets can be compared with word embedding vectors, considering word vectors as another "feature set", the correlation between such vectors and word embeddings (for instance, on predicting k nearest neighbors to a certain word) can report the quality of a DSM [130, 10].

We are aware only of two publicly available English eye movement datasets that one could use in their experiments. The first is the **Provo Corpus** [99] which consists of data of reading 55 paragraphs from 84 native speakers. This dataset could be converted into a list of 1 185 words each of which is associated with a set of 26 eye movement features.

The second dataset is the **Ghent Eye-Tracking Corpus (GECO)** [36] containing data of reading 5 000 sentences from monolingual and bilingual English speakers (33 participants overall). After converting one could obtain a dataset of 987 words, each associated with 48 features.

4 Experimental Data-Driven Evaluation Methods

4.1 QVEC

Building the inverted index of a collection of documents in which each is responsible for a certain category of human knowledge like super-senses in WordNet (e.g. *food*, *animal*, etc.), we can construct the so-called "thesaurus vectors" and use them a proxy for evaluating word embeddings.

The dimensionality of these "thesaurus vectors" is the size of the document collection, and each component in these vectors reports the number of occurrences of the word in a certain document.

For the sake of computational efficiency (to process the large collections), we can also map one component of an embeddings vector to multiple components of thesaurus vectors (or vice versa if the collection is too small) [132].

In the original paper presenting this method, the authors used a so-called "conceptual thesaurus" based on WordNet, but we believe that a set of documents that claims to contain a comprehensive set of the knowledge categories can be used to obtain the "thesaurus vectors".

For instance, *Wikipedia*, which was already similarly used for document vectorization, referring to a method of *Explicit Semantic Analysis*

[56] which was considered for the task of cross-language information retrieval.

4.2 Dictionary Definition Graph

Co-occurrences of words in dictionary definitions could carry information about their relationships [2], we construct a digraph from the set of dictionaries where the nodes are represented by the words, and the values of the edges connecting the word a to the word b are represented by the number of all occurrences of the word b in all the definitions of the word a.

Transforming this graph to a matrix, we obtain a "dictionary vector" for each word, and use these vectors as a proxy of evaluation. Alternatively, one can represent the edges not with simply frequencies of the co-occurrences but the amounts of time when b was encountered as a head in the dependency syntax tree (such an idea can help to identify similarities based on phrases like *a cat is an animal*).

4.3 Cross-Match Test

A Cross-match test is a technique of finding similarity between two high-dimensional sets used to compare blood samples in medicine, and we can use this method for evaluating word embeddings as well.

Determining whether the two sets of values are sampled from the same distribution, we measure the statistical significance of a model, if the correlation of a sample of vectors of two different word vector models is low, then the two compared models probably use different features of the corpus, so it is probably a good result [67].

4.4 Semantic Difference

Characterizing words of the distinctive features (*attributes*), we consider each word in a pair associated with a certain set of attributes. The distance between words is calculated as the difference between the Cartesian product multiplied by the attributes of the word vectors, we can select a pair of attributes of the same category for each pair of non-abstract words (e.g. the category could be *size*, and the distinctive attributed could be *big* and *small*) [90].

There is a certain amount of databases where words are associated with sets of different attributes. One of the examples of such bases is a previously mentioned *BLESS* dataset, which contains 200 pairs of words (for example, for the [*motorcycle*, *moped*] word pair these are the two sets of attributes: [*large*, *small*] and [*fast*, *slow*]) [16].

Another example is *Feature Norms Dataset* containing 24 963 pairs of words, for which a least one pair of distinctive features is selected (for example, for the pair [*airplane*, *helicopter*] the *existence of wings* is selected) [90].

4.5 Semantic Networks

In manually constructed knowledge graphs like WordNet [74], *semantic networks*, the words are organized following their semantic distinctive features based on judgments of the linguists. These graphs provide a measure of similarity for word pairs based on the shortest path in a graph, so such similarity measure can be used as a proxy for the similarity measure of the same pair calculated by word embeddings to evaluate its quality [3].

4.6 Phonosemantic Analysis

The general (and very heuristic) intuition is that the form of a linguistic sign is not arbitrary and has a relationship to its meaning. If that is true, we can use phonosemantic patterns of the word (its phonemes or characters) as a proxy for its meaning. To calculate the phonosemantic difference between two words, one could measure using Levenshtein distance measure, and such metric could be used as a gold standard for evaluation [68]. Notably, this observation was confirmed not only for the Latin alphabet but also for Cyrillic [91].

4.7 Bi-Gram Co-Occurrence Frequency

The distance between the words vectors representing words of a phrase group (e.g. *noun + adjective*) should correlate with the frequency of this group in a corpus (bi-gram co-occurrence frequency). In other words, bi-gram co-occurrence frequency could be used as a gold standard [89].

4.8 Image-Based Evaluation

The method relies on image vectors for word embeddings evaluation and explores whether the similarity spaces generated by two disparate algorithms give rise to similar similarities among high-frequency items [5].

4.9 Closed Domain Evaluation

The method aims at the evaluation of word (and sentence) embeddings from specialized corpora in concept-focused domains [111]; the authors suggest using so-called “ground truths” as a proxy for evaluation [19].

Based on a QUINE corpus, the evaluation consists of a semi-formal definition of the relations of some key terms to other terms, and by defining these interrelations between terms in the corpus, the expert knowledge of the meaning of a term within the corpus is reflected by how the term relates to other terms.

In the case of our Neo-Latin corpus, the domain expert identified that definition (definition) and axioma (axiom) are functional synonyms of principium (principle). Similar to the task discussed above, to successfully complete this task, the cosine distance of the vector of a given target term has to be nearer to the vectors of their functional synonyms than alternative terms.

In the case of principium, definitio and axioma, the cosine distance of the vectors of these terms is expected to be nearer to each other than to other terms. Such a conceptual evaluation grounded in expert knowledge provides a method to evaluate word embeddings intrinsically and, thereby, the quality of their consistency [22].

4.10 Consistency Evaluation

The model is considered consistent if its output does not vary when its input should not trigger variation (i.e., because it is sampled from the same text). Thus, a model can only be as consistent as the input data it is trained on and it requires the experimenter to compute data consistency in addition to vector space consistency.

To evaluate data consistency, we create vectors for target terms in a domain corpus under two conditions: a) random sampling; b) equal split. The “equal split” condition simply corresponds to splitting the data in the middle, thus obtaining two subcorpora of equal size and in diachronic order. Given a pre-trained background space kept frozen across experiments, the vector representation of a target is generated by simple vector addition over its context words.

Therefore, the obtained vector directly represents the context the target term occurs in, and consequently, similar representations (in terms of cosine similarity) mean that the target term is used similarly in different parts of a book/corpus, and is thus consistently learnable. Crucially, though, this measure may interact with data size [21]. A similar metric considered as “reliability“ was checked in [75].

5 Conclusion

The core concern of lexical typology, i.e., how languages express meanings by words, can be approached from slightly different perspectives. We can start from the meanings, or concepts, and ask how these are expressed in different languages, among other things, how semantic domains are distributed among the lexical items across languages.

Lexico-typological research can also start from the expressions (lexemes) and ask what different meanings can be expressed by them or by lexemes that are related to them synchronically and/or diachronically. In this survey we systematized the existing attempts to answer a question of what is a good distributional semantic model and we highlighted that this question always implicitly supposes a question of what is a good model

of the lexicon, and therefore another question of what lexicon is.

We tried to make this paper useful for engineers from the industry as well as linguistics from academia, so we extensively described both well-known "empirical" evaluation methods (such as word similarity task and word analogy task) and experimental methods based on the use of thesauri, semantic networks, or even neuroimaging data.

To not extend this paper to a monstrous size, we tried to focus on overviewing and discussing only those works which were dedicated to both a) "traditional" distributional semantic models aiming to produce representations of lexical units (e.g. Word2Vec), b) the problem of evaluation of the quality of such models or the representations produced by such models, nevertheless what was meant by "quality" by the authors of such works.

What we believe is that the notion of the quality of distributional word representations is heavily tied to the notion of their linguistic representativeness, i.e., the degree of being a proper model of a lexicon. Despite this notion of representativeness being grounded to theoretical linguistics and the legacy of formal analysis of semantics, we have shown in Section 3 dedicated to the so-called subconscious evaluation methods that the exploration of cognitive dynamics could be a promising direction towards understanding mechanisms of distributional semantics.

Another view on the quality of the DSMs that we consider important for further studies (despite it has not been included in this survey) considers the reliability of distributional semantic models from the position of fairness and prejudices [11] a representative model should not contain prejudices against certain groups of people (by their gender, ethnicity, sexual orientation, etc.).

If we accept this assumption, we should understand how underlying mechanisms of such fairness bias look like, and how to automatically remove bias from vector spaces. We consider that this fairness property is also an important feature of distributional semantics, and experiments grounding either to DSM's quality, performance, reliability, linguistic motivation, or whatnot, also should have such issue in mind.

As we mentioned in the first section of the paper, the trends in NLP are now heavily shifted towards sentence embeddings, and traditional word vector representations became a "niche" topic. However, we assume that algorithms like ELMo [120], BERT [42] and GPT-2 [121] do not provide so much of scientific interest from the position of the lexicon as traditional word embeddings, as they are not relying on single words, but adopt a more "syntactically savvy" notion of linguistic contexts, in which word semantics are reconstructed by specific syntactic configurations.

There is a hypothesis that such evaluation of representation in context is more reliable, and experiments on such context-based models would be more representative. But context-based setting just grounds to one of the semantic views which are not assumed to be absolutely correct [58].

The theory that lexical semantics is not grounded to the context [81] gains motivation from cognitive studies. It basically gives the main attention to how words obtain meaning in human cognition and interact with other linguistic units, while the context-sensitive approach is not compatible with the idea that syntax a priori acts as the scaffolding that guides distributional analysis. This survey could be considered as a small step forward to bigger planned research of computational formalisms for lexical semantics.

We plan to give more attention to other semantic theories and their theoretical background to propose a more detailed exploration of distributional semantics, for example, the referential one. One of the possible directions of further work on the topic of this survey goes to the intersection of referential and distributional theory. Taking roots from structural theory, certain theories try to ground distributional hypothesis to an interpretable framework [26].

Probably one of the most notorious works in this field goes to **Compositional Distributional Semantic** framework [34], DisCoCat, which suggests construction representations of sentences or documents not through arithmetic operations on word vectors, but by categorical logical operators.

Recent studies propose its experimental support [66] and we assume that such approach could be

more efficient than common distributional theory on certain downstream tasks like anaphora and ellipsis resolution [142].

We hope that the work done by writing this survey at least could be helpful to scholars to look at distributional semantics from a new scope and start to treat word vectors not only as black-box tools for resolving downstream tasks but as linguistic formalisms that have their benefits and limitations from the language perspective.

Despite most of the experimental studies on a similar topic doubting the generalizing ability of distributional semantics, we do not suggest refusing it! In opposite, we argue that we should give more attention to its detailed investigation. But as a matter of fact, we should not narrow the computational semantics research to this theory.

As we know the limitations of this theory, we can draw ideas from other semantics theories to overcome them. Maybe it is time to shake off the dust from abandoned semantics theories and revise their ideas since feasibly the forgotten evening/morning star is the one that leads us to clarity.

Acknowledgments

The reported study was funded by the Russian Foundation for Basic Research project 20-37-90153 “Development of framework for distributional semantic models evaluation”.

References

1. **Abnar, S., Ahmed, R., Mijnheer, M., Zuidema, W. (2018).** Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018), pp. 57–66.
2. **Acs, J., Kornai, A. (2016).** Evaluating embeddings on dictionary-based similarity. Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, pp. 78–82.
3. **Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., Soroa, A. (2009).** A study on similarity and relatedness using distributional and wordnet-based approaches. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 19–27.
4. **Almuhareb, A. (2006).** Attributes in lexical acquisition. Ph.D. thesis, University of Essex.
5. **Amatuni, A., He, E., Bergelson, E. (2018).** Preserved structure across vector space representations. arXiv preprint arXiv:1802.00840.
6. **Antoniak, M., Mimno, D. (2018).** Evaluating the stability of embedding-based word similarities. Transactions of the Association for Computational Linguistics, Vol. 6, pp. 107–119.
7. **Artemova, E., Bakarov, A., Artemov, A., Burnaev, E., Sharaev, M. (2020).** Data-driven models and computational tools for neurolinguistics: a language technology perspective. Journal of Cognitive Science, Vol. 21, No. 1, pp. 15–52.
8. **Asr, F. T., Zinkov, R., Jones, M. (2018).** Querying word embeddings for similarity and relatedness. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1, pp. 675–684.
9. **Auguste, J., Rey, A., Favre, B. (2017).** Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks. Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP, pp. 21–26.
10. **Bakarov, A. (2018).** Can eye movement data be used as ground truth for word embeddings evaluation? arXiv preprint arXiv:1804.08749.
11. **Bakarov, A. (2020).** Did you just assume my vector? Detecting gender stereotypes in word embeddings. International Conference on Analysis of Images, Social Networks and Texts, Springer, pp. 3–10.
12. **Baker, S., Reichart, R., Korhonen, A. (2014).** An unsupervised model for instance level subcategorization acquisition. EMNLP, pp. 278–289.

13. **Baroni, M., Dinu, G., Kruszewski, G. (2014).** Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 238–247.

14. **Baroni, M., Evert, S., Lenci, A. (2008).** ESSLLI 2008 workshop on distributional lexical semantics. Hamburg, Germany: Association for Logic, Language and Information.

15. **Baroni, M., Lenci, A. (2010).** Distributional memory: A general framework for corpus-based semantics. Computational Linguistics, Vol. 36, No. 4, pp. 673–721.

16. **Baroni, M., Lenci, A. (2011).** How we BLESSed distributional semantic evaluation. Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, Association for Computational Linguistics, pp. 1–10.

17. **Baroni, M., Murphy, B., Barbu, E., Poesio, M. (2010).** Strudel: A corpus-based semantic model based on properties and types. Cognitive science, Vol. 34, No. 2, pp. 222–254.

18. **Batchkarov, M., Kober, T., Reffin, J., Weeds, J., Weir, D. (2016).** A critique of word similarity as a method for evaluating distributional semantic models. Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, Association for Computational Linguistics, pp. 7–12.

19. **Betti, A., Reynaert, M., Ossenkoppele, T., Oortwijn, Y., Salway, A., Bloem, J. (2020).** Expert concept-modeling ground truth construction for word embeddings evaluation in concept-focused domains. Proceedings of the 28th International Conference on Computational Linguistics, pp. 6690–6702.

20. **Blair, P., Merhav, Y., Barry, J. (2016).** Automated generation of multilingual clusters for the evaluation of distributed representations. arXiv preprint arXiv:1611.01547.

21. **Bloem, J., Fokkens, A., Herbelot, A. (2019).** Evaluating the consistency of word embeddings from small data. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp. 132–141.

22. **Bloem, J., Parisi, M. C., Reynaert, M., Oortwijn, Y., Betti, A. (2020).** Distributional semantics for neo-Latin. Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages, pp. 84–93.

23. **Bloomfield, L. (1914).** An introduction to the study of language. H. Holt.

24. **Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2016).** Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.

25. **Boleda, G., Erk, K. (2015).** Distributional semantic features as semantic primitives - or not. 2015 AAAI Spring Symposium Series.

26. **Boleda, G., Herbelot, A. (2016).** Formal distributional semantics: Introduction to the special issue. Computational Linguistics, Vol. 42, No. 4, pp. 619–635.

27. **Bruni, E., Tran, N. K., Baroni, M. (2014).** Multimodal distributional semantics. J. Artif. Intell. Res.(JAIR), Vol. 49, No. 2014, pp. 1–47.

28. Bybee, J. L., Perkins, R. D., Pagliuca, W., others **(1994).** The evolution of grammar: Tense, aspect, and modality in the languages of the world, volume 196. University of Chicago Press Chicago.

29. **Camacho-Collados, J., Navigli, R. (2016).** Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations. ACL Workshop on Evaluating Vector Space Representations for NLP, pp. 43–50.

30. **Camacho-Collados, J., Pilehvar, M. T., Collier, N., Navigli, R. (2017).** Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017). Vancouver, Canada.

31. **Casacuberta, S., Halevy, K., Blasi, D. E. (2021).** Evaluating word embeddings with categorical modularity. arXiv preprint arXiv:2106.00877.

32. **Che, X., Ring, N., Raschkowski, W., Yang, H., Meinel, C. (2017).** Traversal-free word vector evaluation in analogy space. Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP, pp. 11–15.

33. **Chiu, B., Korhonen, A., Pyysalo, S. (2016).** Intrinsic evaluation of word vectors fails to predict extrinsic performance. Proceedings of

the 1st Workshop on Evaluating Vector Space Representations for NLP, pp. 1–6.

34. **Coecke, B., Sadrzadeh, M., Clark, S. (2010).** Mathematical foundations for a compositional distributed model of meaning. Lambek Festschirft, Linguistic Analysis, Vol. 36.
35. **Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P. (2011).** Natural language processing (almost) from scratch. Journal of Machine Learning Research, Vol. 12, No. Aug, pp. 2493–2537.
36. **Cop, U., Dirix, N., Drieghe, D., Duyck, W. (2017).** Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. Behavior research methods, Vol. 49, No. 2, pp. 602–615.
37. **Cristofaro, S. (2010).** Semantic maps and mental representation. Linguistic Discovery, Vol. 8, No. 1.
38. **Croft, W. (2002).** Typology and universals. Cambridge University Press.
39. **Croft, W., Poole, K. T. (2008).** Inferring universals from grammatical variation: Multidimensional scaling for typological analysis.
40. **Cruse, A. (2010).** Meaning in language: An introduction to semantics and pragmatics.
41. **Cysouw, M., Haspelmath, M., Malchukov, A. (2010).** Introduction to the special issue "Semantic maps: Methods and applications". Linguistic Discovery, Vol. 8, No. 1, pp. 1–3.
42. **Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018).** Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
43. **Duncan, K. J., Pattamadilok, C., Knierim, I., Devlin, J. T. (2009).** Consistency and variability in functional localisers. Neuroimage, Vol. 46, No. 4, pp. 1018–1026.
44. **Erk, K. (2016).** What do you know about an alligator when you know the company it keeps? Semantics and Pragmatics, Vol. 9, pp. 17–1.
45. **Ettinger, A., Feldman, N. H., Resnik, P., Phillips, C. (2016).** Modeling N400 amplitude using vector space models of word representation. Proceedings of the 38th annual conference of the Cognitive Science Society, pp. 1445–1450.
46. **Ettinger, A., Linzen, T. (2016).** Evaluating vector space models using human semantic priming results. Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, pp. 72–77.
47. **Faruqui, M., Tsvetkov, Y., Rastogi, P., Dyer, C. (2016).** Problems with evaluation of word embeddings using word similarity tasks. arXiv preprint arXiv:1605.02276.
48. **Fellbaum, C. (2010).** WordNet. In Theory and applications of ontology: computer applications. Springer, pp. 231–243.
49. **Ferretti, T. R., McRae, K., Hatherell, A. (2001).** Integrating verbs, situation schemas, and thematic role concepts. Journal of Memory and Language, Vol. 44, No. 4, pp. 516–547.
50. **Fillmore, C. J. (1976).** Frame semantics and the nature of language. Annals of the New York Academy of Sciences, Vol. 280, No. 1, pp. 20–32.
51. **Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E. (2001).** Placing search in context: The concept revisited. Proceedings of the 10th international conference on World Wide Web, ACM, pp. 406–414.
52. **Finley, G., Farmer, S., Pakhomov, S. (2017).** What analogies reveal about word vectors and their compositionality. Proceedings of the 6th joint conference on lexical and computational semantics (* SEM 2017), pp. 1–11.
53. **Firth, J. R. (1957).** A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis.
54. **Friedman, H. H., Amoo, T. (1999).** Rating the rating scales. Friedman, Hershey H. and Amoo, Taiwo (1999)." Rating the Rating Scales." Journal of Marketing Management, Winter, pp. 114–123.
55. **Fujinuma, Y., Boyd-Graber, J., Paul, M. (2019).** A resource-free evaluation metric for cross-lingual word embeddings based on graph modularity. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4952–4962.
56. **Gabrilovich, E., Markovitch, S. (2007).** Computing semantic relatedness using wikipedia-based explicit semantic analysis. IJcAI, volume 7, pp. 1606–1611.

57. **Gao, B., Bian, J., Liu, T.-Y. (2014).** Wordrep: A benchmark for research on learning word representations. arXiv preprint arXiv:1407.1640.

58. **Geeraerts, D. (2010).** Theories of lexical semantics. Oxford University Press.

59. **Gerz, D., Vulić, I., Hill, F., Reichart, R., Korhonen, A. (2016).** Simverb-3500: A large-scale evaluation set of verb similarity. arXiv preprint arXiv:1608.00869.

60. **Gittens, A., Achlioptas, D., Mahoney, M. W. (2017).** Skip-gram- zipf+ uniform= vector additivity. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 69–76.

61. **Gladkova, A., Drozd, A. (2016).** Intrinsic evaluations of word embeddings: What can we do better? Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, pp. 36–42.

62. **Gladkova, A., Drozd, A., Matsuoka, S. (2016).** Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. Proceedings of the NAACL Student Research Workshop, pp. 8–15.

63. **Glynn, D., Fischer, K. (2010).** Quantitative methods in cognitive semantics: Corpus-driven approaches, volume 46. Walter de Gruyter.

64. **Goldsmith, J. A. (2005).** The legacy of Zellig Harris: Language and information into the 21st century, vol. 1: Philosophy of science, syntax and semantics. Language, Vol. 81, No. 3, pp. 719–736.

65. **Greenberg, C., Demberg, V., Sayeed, A. (2015).** Verb polysemy and frequency effects in thematic fit modeling. Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics. Association for Computational Linguistics, Denver, Colorado, pp. 48–57.

66. **Grefenstette, E., Sadrzadeh, M. (2011).** Experimental support for a categorical compositional distributional model of meaning. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 1394–1404.

67. **Gurnani, N. (2017).** Hypothesis testing based intrinsic evaluation of word embeddings. arXiv preprint arXiv:1709.00831.

68. **Gutiérrez, E. D., Levy, R., Bergen, B. (2016).** Finding non-arbitrary form-meaning systematicity using string-metric learning for kernel regression. ACL (1).

69. **Halawi, G., Dror, G., Gabrilovich, E., Koren, Y. (2012).** Large-scale learning of word relatedness with constraints. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 1406–1414.

70. **Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., Zinke, W., Stadler, J. (2014).** A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. Scientific data, Vol. 1, pp. 140003.

71. **Harris, Z. S. (1954).** Distributional structure. Word, Vol. 10, No. 2-3, pp. 146–162.

72. **Hashimoto, T. B., Alvarez-Melis, D., Jaakkola, T. S. (2016).** Word embeddings as metric recovery in semantic spaces. Transactions of the Association for Computational Linguistics, Vol. 4, pp. 273–286.

73. **Haspelmath, M. (2003).** The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In The new psychology of language. Psychology Press, pp. 217–248.

74. **Hearst, M. (1998).** Wordnet: An electronic lexical database and some of its applications.

75. **Hellrich, J., Hahn, U. (2016).** Bad company - neighborhoods in neural embedding spaces considered harmful. Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers, pp. 2785–2796.

76. **Herdağdelen, A., Erk, K., Baroni, M. (2009).** Measuring semantic relatedness with vector space models and random walks. Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing, Association for Computational Linguistics, pp. 50–53.

77. **Hill, F., Reichart, R., Korhonen, A. (2016).** Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics.

78. **Hollenstein, N., Van der Lek, A., Zhang, C. (2020).** Cognival in action: An interface for customizable cognitive word embedding

evaluation. Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations, pp. 34–40.

79. **Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., Yap, M. J., Bengson, J. J., Niemeyer, D., Buchanan, E. (2013).** The semantic priming project. Behavior Research Methods, Vol. 45, No. 4, pp. 1099–1114.
80. **Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., Gallant, J. L. (2016).** Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, Vol. 532, No. 7600, pp. 453–458.
81. **Jackendoff, R. (1976).** Toward an explanatory semantic representation. Linguistic inquiry, Vol. 7, No. 1, pp. 89–150.
82. **Jarmasz, M., Szpakowicz, S. (2004).** Roget's thesaurus and semantic similarity. Recent Advances in Natural Language Processing III: Selected Papers from RANLP, Vol. 2003, pp. 111.
83. **Jastrzebski, S., Leśniak, D., Czarnecki, W. M. (2017).** How to evaluate word embeddings? On importance of data efficiency and simple supervised tasks. arXiv preprint arXiv:1702.02170.
84. **Jones, M. N., Kintsch, W., Mewhort, D. J. (2006).** High-dimensional semantic space accounts of priming. Journal of memory and language, Vol. 55, No. 4, pp. 534–552.
85. **Joseph, K., Carley, K. M. (2016).** Relating semantic similarity and semantic association to how humans label other people. Proceedings of the First Workshop on NLP and Computational Social Science, pp. 1–10.
86. **Jurgens, D. A., Turney, P. D., Mohammad, S. M., Holyoak, K. J. (2012).** Semeval-2012 task 2: Measuring degrees of relational similarity. Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, Association for Computational Linguistics, pp. 356–364.
87. **Kiela, D., Hill, F., Clark, S. (2015).** Specializing word embeddings for similarity or relatedness. EMNLP, pp. 2044–2048.
88. **Koptjevskaja-Tamm, M., Rakhilina, E., Vanhove, M. (2015).** The semantics of lexical typology. The Routledge Handbook of Semantics, pp. 434.
89. **Kornai, A., Kracht, M. (2015).** Lexical semantics and model theory: Together at last? Proceedings of the 14th Meeting on the Mathematics of Language (MoL 2015), pp. 51–61.
90. **Krebs, A., Paperno, D. (2016).** Capturing discriminative attributes in a distributional space: Task proposal. ACL 2016, pp. 51.
91. **Kutuzov, A. (2017).** ArbitrAriness of linguistic sign questioned: correlation between word form and meaning in Russian. Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii, Vol. 1, No. 16 (23), pp. 109–120.
92. **Landauer, T. K., Dumais, S. T. (1997).** A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, Vol. 104, No. 2, pp. 211.
93. **Langacker, R. W. (1987).** Foundations of cognitive grammar: Theoretical prerequisites, volume 1. Stanford university press.
94. **Lapesa, G., Evert, S. (2013).** Evaluating neighbor rank and distance measures as predictors of semantic priming. Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2013), pp. 66–74.
95. **Lehmann, C. (1990).** Towards lexical typology. Studies in typology and diachrony: Papers presented to Joseph H. Greenberg on his 75th birthday, volume 161, pp. 185.
96. **Lenci, A. (2008).** Distributional semantics in linguistic and cognitive research. Italian journal of linguistics, Vol. 20, No. 1, pp. 1–31.
97. **Levy, O., Goldberg, Y. (2014).** Linguistic regularities in sparse and explicit word representations. Proceedings of the eighteenth conference on computational natural language learning, pp. 171–180.
98. **Liza, F. F., Grzes, M. (2016).** An improved crowdsourcing based evaluation technique for word embedding methods. ACL 2016, pp. 55.
99. **Luke, S. G., Christianson, K. (2017).** The Provo corpus: A large eye-tracking corpus with

predictability norms. Behavior Research Methods, pp. 1–8.

100. **Luong, T., Socher, R., Manning, C. D. (2013).** Better word representations with recursive neural networks for morphology. CoNLL, pp. 104–113.
101. **Mandera, P., Keuleers, E., Brysbaert, M. (2017).** Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. Journal of Memory and Language, Vol. 92, pp. 57–78.
102. **McDonald, S., Brew, C. (2004).** A distributional model of semantic context effects in lexical processing. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp. 17.
103. **McRae, K., Ferretti, Liane Amyote, T. R. (1997).** Thematic roles as verb-specific concepts. Language and cognitive processes, Vol. 12, No. 2-3, pp. 137–176.
104. **McRae, K., Spivey-Knowlton, M. J., Tanenhaus, M. K. (1998).** Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. Journal of Memory and Language, Vol. 38, No. 3, pp. 283–312.
105. **Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013).** Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
106. **Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. (2013).** Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, pp. 3111–3119.
107. **Mikolov, T., Yih, W.-t., Zweig, G. (2013).** Linguistic regularities in continuous space word representations. hlt-Naacl, volume 13, pp. 746–751.
108. **Milajevs, D., Griffiths, S. (2016).** A proposal for linguistic similarity datasets based on commonality lists. arXiv preprint arXiv:1605.04553.
109. **Miller, G. A., Charles, W. G. (1991).** Contextual correlates of semantic similarity. Language and cognitive processes, Vol. 6, No. 1, pp. 1–28.
110. **Nissim, M., van Noord, R., van der Goot, R. (2019).** Fair is better than sensational: Man is to doctor as woman is to doctor. arXiv preprint arXiv:1905.09866.
111. **Nooralahzadeh, F., Øvrelid, L., Lønning, J. T. (2018).** Evaluation of domain-specific word embeddings using knowledge resources. Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).
112. **Osgood, C. E., Suci, G. J., Tannenbaum, P. H. (1978).** The measurement of meaning. 1957. Urbana: University of Illinois Press.
113. **Padó, S., Lapata, M. (2007).** Dependency-based construction of semantic space models. Computational Linguistics, Vol. 33, No. 2, pp. 161–199.
114. **Padó, U. (2007).** The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing.
115. **Panchenko, A., others (2013).** Similarity measures for semantic relation extraction. Ph.D. thesis, PhD thesis, Université catholique de Louvain & Bauman Moscow State Technical University.
116. **Parviz, M., Johnson, M., Johnson, B., Brock, J. (2011).** Using language models and latent semantic analysis to characterise the N400m neural response. Proceedings of the Australasian Language Technology Association Workshop 2011, pp. 38–46.
117. **Peinelt, N., Liakata, M., Nguyen, D. (2019).** Aiming beyond the obvious: Identifying non-obvious cases in semantic similarity datasets. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2792–2798.
118. **Pennington, J., Socher, R., Manning, C. (2014).** Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543.
119. **Pereira, F., Gershman, S., Ritter, S., Botvinick, M. (2016).** A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. Cognitive neuropsychology, Vol. 33, No. 3-4, pp. 175–190.
120. **Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. (2018).** Deep

contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 2227–2237.

121. **Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019).** Language models are unsupervised multitask learners. OpenAI Blog, Vol. 1, pp. 8.
122. **Radinsky, K., Agichtein, E., Gabrilovich, E., Markovitch, S. (2011).** A word at a time: computing word relatedness using temporal semantic analysis. Proceedings of the 20th international conference on World wide web, ACM, pp. 337–346.
123. **Resnik, P. (1995).** Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007.
124. **Rubenstein, H., Goodenough, J. B. (1965).** Contextual correlates of synonymy. Communications of the ACM, Vol. 8, No. 10, pp. 627–633.
125. **Sahlgren, M. (2008).** The distributional hypothesis. Italian Journal of Disability Studies, Vol. 20, pp. 33–53.
126. **Salicchi, L., Lenci, A., Chersoni, E. (2021).** Looking for a role for word embeddings in eye-tracking features prediction: Does semantic similarity help? Proceedings of the 14th International Conference on Computational Semantics (IWCS), pp. 87–92.
127. **Sayeed, A., Greenberg, C., Demberg, V. (2016).** Thematic fit evaluation: an aspect of selectional preferences. ACL 2016, pp. 99.
128. **Schnabel, T., Labutov, I., Mimno, D. M., Joachims, T. (2015).** Evaluation methods for unsupervised word embeddings. EMNLP, pp. 298–307.
129. **Schwartz, D., Mitchell, T. (2019).** Understanding language-elicited EEG data by predicting it from a fine-tuned language model. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 43–57.
130. **Søgaard, A. (2016).** Evaluating word embeddings with fMRI and eye-tracking. ACL 2016, pp. 116.
131. **Szymanski, T. (2017).** Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers), pp. 448–453.
132. **Tsvetkov, Y., Faruqui, M., Ling, W., Lample, G., Dyer, C. (2015).** Evaluation of word vector representations by subspace alignment. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2049–2054.
133. **Turian, J., Ratinov, L., Bengio, Y. (2010).** Word representations: a simple and general method for semi-supervised learning. Proceedings of the 48th annual meeting of the association for computational linguistics, Association for Computational Linguistics, pp. 384–394.
134. **Turney, P. (2001).** Mining the web for synonyms: PMI-IR versus LSA on TOEFL. Machine Learning: ECML 2001, pp. 491–502.
135. **Turney, P. D. (2008).** The latent relation mapping engine: Algorithm and experiments. Journal of Artificial Intelligence Research, Vol. 33, pp. 615–655.
136. **Turney, P. D., Littman, M. L., Bigham, J., Shnayder, V. (2003).** Combining independent modules to solve multiple-choice synonym and analogy problems. arXiv preprint cs/0309035.
137. **Turney, P. D., Pantel, P. (2010).** From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, Vol. 37, pp. 141–188.
138. **Utsumi, A. (2020).** Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. Cognitive Science, Vol. 44, No. 6, pp. e12844.
139. **Vandekerckhove, B., Sandra, D., Daelemans, W. (2009).** A robust and extensible exemplar-based model of thematic fit. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 826–834.
140. **Vylomova, E., Rimell, L., Cohn, T., Baldwin, T. (2015).** Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. arXiv preprint arXiv:1509.01692.

141. Wälchli, B. (2010). Similarity semantics and building probabilistic semantic maps from parallel texts. Linguistic Discovery, Vol. 8, No. 1, pp. 331–371.

142. Wijnholds, G., Sadrzadeh, M. (2019). A typedriven vector semantics for ellipsis with anaphora using lambek calculus with limited contraction. arXiv preprint arXiv:1905.01647.

143. Wittgenstein, L. (2010). Philosophical investigations. John Wiley & Sons.

144. Yang, D., Powers, D. M. (2006). Verb similarity on the taxonomy of wordnet. The Third International WordNet Conference: GWC 2006, Masaryk University.

Article received on 07/04/2022; accepted on 20/07/2022.
Corresponding author is Amir Bakarov.