In search of semantic distance: metaphorical and non-metaphorical constructions in static and contextual embeddings

Mojca Brglez[1 2] *and Špela Vintar*[1 2]
[1] University of Ljubljana
[2] "Jožef Stefan" Institute

ABSTRACT

Phrases such as *burning question, digital waste* or *invasion of technology* are relatively ordinary expressions understood by any speaker of English. While diverse in structure and meaning, they demonstrate a semantic tension between the basic meanings of a metaphoric and a non-metaphoric constituent. Albeit frequent in discourse, they remain a challenge for automatic language processing systems, especially for smaller, less represented languages. In this work, we inspect a broad array of language models to embed noun phrases in Slovene and investigate the potential of word embeddings to identify metaphoric phrases via the semantic distance of its constituents as measured via cosine similarity. The study shows both static and contextual monolingual embeddings encode relevant semantic information while multilingual embeddings demonstrate no significant effect in this experimental setting. Moreover, the study unravels the most effective layers for basic meaning representation and highlights the influence of other, non-semantic factors on cosine similarity. By shedding light on these mechanisms, the study provides new insights for both metaphor processing and our understanding of the inner workings of language models.

Keywords: metaphor, semantic distance, word embeddings, noun phrases

Journal of Language Modelling Vol 13, No 2 (2025), pp. 207–260

1 INTRODUCTION

In the past few decades, metaphor has been recognized not just as a decorative feature of language but also as a powerful cognitive and communicative device (Lakoff and Johnson 1980; Steen 2017; Burgers *et al.* 2016) occurring in all types of discourse (Reijnierse *et al.* 2019; Cameron 2003; Semino 2008). The main underlying mechanism of metaphor involves representing one domain in terms of another (Lakoff and Johnson 1980, 2003; Kövecses 2020). For example, in the expressions *political storm*, *climate in the Congress*, or *dark clouds over his presidential campaign*, the domain of POLITICS is represented through the domain of WEATHER. The represented domain, usually more abstract, complex, and unfamiliar, is called the **target domain**, and the domain it is represented by is called the **source domain**, which is usually more concrete, familiar, and based on physical experience.

By representing one domain through another, metaphor not only equates two things but expresses implicit meanings, as it highlights only some aspects about the target while hiding others. By framing the target through a specific source, metaphors can express emotional attitudes (Goatly 2011), influence reasoning (Thibodeau and Boroditsky 2011), and serve as persuasive tools (Charteris-Black 2004; Boeynaems *et al.* 2017). In linguistics, metaphors are recognized as one of the main drivers of semantic change leading to word polysemy (Blank 1999), making them essential for identifying novel meanings and informing lexicographic work. Metaphor understanding is also crucial – yet non-trivial – in contexts such as second language acquisition (Littlemore and Low 2006) and translation (Massey 2021). Moreover, metaphorical language still presents a challenge in the field of natural language processing, where even the newest generations of language models struggle to correctly detect or interpret metaphors, at least in smaller languages such as Slovene.[1] The challenges of disambiguating

[1] At the time of writing, we experimented with the then-cutting edge model GPT-4o and could still find examples where metaphors get mislabelled or misinterpreted, which might be a consequence of English interference. For example, it mistakenly labels the noun in the phrase *politični besednjak* ‘political vocabulary’ as metaphorical and provides the following explanation: *“Besednjak” (vocabu-*

and understanding metaphorical language have also been shown to influence computational tasks not directly targeting metaphors, such as machine translation (Mao *et al.* 2018; Wang *et al.* 2024) and sentiment analysis (Rentoumi *et al.* 2012), making metaphor processing an important consideration across many fields and disciplines.

Metaphors can be expressed linguistically in many unconstrained ways. First of all, they can take the shape of any part of speech. Even function words such as prepositions can be used metaphorically by extending their literal meaning to a more abstract level. These conceptual mappings are deeply entrenched in our language and cognition, are not usually noticed by native speakers, and hence are not the focus of our study. A more conspicuous type of metaphor showing linguistic creativity is found in content words, especially adjectives and nouns (Do Dinh *et al.* 2018). Secondly, metaphors can materialize as single words or wind through long passages of text. Nevertheless, they can be recognized with the help of two main linguistic cues. First, there is often some **semantic incongruity** between the metaphorically used word and its context. According to Wilks (1978), metaphors can be seen as "**selectional preference violations**" (SPVs): the context of the metaphorically used word is not the context this word is usually associated with (cf. *political storm*, *thunder storm*). On the other hand, metaphorically used expressions do not only show "external" incongruity, that is, a clash with the context they appear in, but also an internal one, within themselves, as they exhibit a type of **polysemy**. That is, the metaphoric meaning of a word in context is different from its **basic meaning**. Cruse (2006) also calls this the "default" or "primary" meaning of words and defines it as "the meaning [that] is intuitively given in the absence of any context" (Cruse 2006, p. 42). The polysemy of metaphorically used words is also the defining factor of the most frequently used procedure for manual metaphor identification in texts (metaphor identification procedure MIP, proposed by

---

lary) is metaphorical because political discourse is not an actual dictionary but a set of commonly used terms and narratives in politics. This explanation would correspond to the English word *dictionary*, while the Slovene *besednjak* 'vocabulary' does not, in any case, denote an actual physical dictionary. The basic and contextual meanings of the Slovene word are identical, i.e., 'a set of commonly used terms'.

Pragglejaz 2007, and its amended successor MIPVU proposed by Steen 2010), where annotators look for a discrepancy between the contextual meaning and the basic meaning.

The two aforementioned linguistic properties, either independently or in combination, have often also guided automatic metaphor processing. Older approaches to metaphor modelling used a diverse set of resources and features, from distributional vectors such as those created with the help of latent semantic analysis (e.g., Kintsch 2000; Utsumi 2011), linguistic resources such as WordNet (e.g., Krishnakumaran and Zhu 2007), conceptual features such as concreteness (e.g., Turney *et al.* 2011), and word embeddings obtained through deep-learning (e.g., Su *et al.* 2017; Mao *et al.* 2018). State-of-the-art approaches use task-specific neural models (e.g., Choi *et al.* 2021; Babieno *et al.* 2022; Wang *et al.* 2023), while some recent studies have also tried to leverage the power of generative models (e.g., Wachowiak and Gromann 2023; Liang *et al.* 2024). However, these still lack in performance, and, more importantly, the inner mechanisms of such models are difficult to interpret. As a consequence, their erroneous predictions are also impossible to mitigate without additional, usually external, resources.

Metaphor identification approaches also differ depending on the level of metaphor processing: word, syntactic relation, or sentence. On the word-level, the task is to determine the metaphoricity of a single (or each) word. On the sentence-level, the whole sentence is classified as either containing metaphor(s) or not. In this study, we investigate metaphoricity on the **relation-level** wherein the unit of analysis is a pair of words connected by a syntactic relation, e.g., verb-object (***break** a promise*) or adjective-noun constructions (***deep** thought*). Related to and sometimes overlapping with the task of relation-level metaphor identification is the task of identifying multi-word expressions (MWEs). The class of MWEs includes phraseological units such as idioms and proverbs, as well as other fixed expressions such as compounds and collocations (Gantar *et al.* 2018). However, the overlap mostly concerns the more conventionalized, lexicalized metaphoric expressions and does not cover novel, less fixed metaphoric formulations. Another related task, more connected to word-level metaphoricity detection, is word sense disambiguation (WSD). Here, the goal is to identify the correct meaning of a polysemous lexeme in context and classify it into one of the existing dictionary senses. Senses of polysemous words are often established through repetitive metaphoric or metonymic extensions (Cruse 2000, p. 112). For example, the word *chicken* can be used in (at least) three senses: "animal" (basic meaning); "meat of the animal" (metonymic extension), "cowardly person" (metaphoric extension).

The large majority of research work in automatic metaphor processing was conducted for the English language, while far fewer studies have investigated methods for less represented languages. This is also the case for Slovene, where computational metaphor processing has only emerged in recent years and has benefitted from only a few studies (Brglez *et al.* 2021; Zwitter Vitez *et al.* 2022; Klemen and Robnik-Šikonja 2023; Brglez 2023; Brglez *et al.* 2025). Among these, Brglez 2023 is the only approach to date testing the direct use of word embeddings.

In this study, we aim to determine whether the identification of relation-level metaphors is possible, or at least facilitated, by the dissimilarity between the basic meanings of the relation constituents. We compare various static and contextual embeddings, and explore the layers and strategies that are most suitable for this purpose. By examining the representation of basic word meanings, this study helps narrow the gap in metaphor processing in Slovene and contributes to the broader field of investigating the "black box" of neural language models.

The contributions of our study are the following:

1. We extend previous studies of metaphor in Slovene, based on biased datasets, to a dataset extracted from a large-scale corpus, which is annotated based on linguistic principles and contains 'real-world' and varied data;
2. We apply a relation-level rather than word- or sentence-level approach to metaphor modelling in Slovene, for the first time on such a large dataset;
3. We include a wide array of Slovene and multilingual neural embedding models to create word representations and provide layerwise analyses of contextual models;
4. We perform a manual qualitative analysis to correlate the cosine similarity metric, which was extensively used in a variety of

studies, with both semantic similarity as well as other linguistic and distributional-semantic factors;

5. We conduct experiments separately for constructions involving adjective-noun (amod) and noun-noun (nmod) relations, showing a non-negligible effect of syntactic structure and word class on cosine similarity.

2 RELATED WORK

The rise in attention to metaphor in general as well as to metaphor identification in discourse has its roots in the ideas of conceptual metaphor theory (CMT, Lakoff and Johnson 1980, 2003), one of the cornerstones of cognitive linguistics. Here, metaphors are recognized as cognitive devices operating not only on the level of language but on the level of conceptual domains: by framing one domain (target) in terms of another domain (source). For automatic metaphor identification, most approaches make use of two main characteristics of metaphors guided by linguistic theory: word polysemy as epitomized by the metaphor identification procedure (MIP) and contextual incongruity through selectional preference violation (SPV).

The state of the art in word-level metaphor identification uses the paradigm of supervised learning, relying on large pre-trained language models fine-tuned on large annotated corpora. In English, the best performing models have been proposed by Choi *et al.* (2021), Lin *et al.* (2021), Elzohbi and Zhao (2024), Li *et al.* (2023b), and Babieno *et al.* (2022). In all of these, the authors try to explicitly exploit the principles of MIP and SPV by presenting the models with input representations based on the target word in different contexts. These approaches achieve performance ranging up to 0.798 in F_1 score on VUA-20 (the largest and most balanced English metaphor dataset); however, Elzohbi and Zhao (2024) also note the performance varies by part of speech.

For Slovene, Klemen and Robnik-Šikonja 2023 is the only word-level metaphor detection approach. They test four pre-trained BERT-based models: monolingual SloBERTA (Ulčar and Robnik-Šikonja 2021), trilingual CroSloEngual BERT (CSE BERT, Ulčar and Robnik-Šikonja 2020), massively multilingual mBERT (Devlin *et al.* 2019), and XLM-RoBERTa (Conneau *et al.* 2020). They also test the multilingual models in both multi-lingual and cross-lingual settings. In the first, they train the models on both Slovene (KOMET, Antloga 2020a; G-KOMET, Antloga and Donaj 2022) and English (VUA, Steen 2010) datasets, and in the second, they train on the English data only and evaluate on Slovene. The highest overall mean F_1 score on KOMET is 0.607 with CSE BERT trained on Slovene and English data. Models within the monolingual and multilingual training categories are comparable; however, cross-lingual models perform much worse. The results per part of speech show that the model is much better at classifying prepositions, the largest class of metaphors in the KOMET dataset, while the performance is much lower for other parts of speech. When comparing models only on the prediction of nouns and verbs, the monolingual SloBERTa performs best.

Compared to other linguistic processing tasks (e.g., part-of-speech tagging), metaphor identification is arguably more difficult. One of the reasons for the still somewhat inferior results is that it became a subject of interest in NLP studies much later. Secondly, although neural approaches have recorded steady improvements over the years, less attention has been paid to the inner workings of those models, making their erroneous outputs hard to interpret or correct. In our current study, we are interested in directly using word embeddings for metaphor identification, that is, without training a specialized model on a labelled dataset in a supervised manner. Because our work is focused on relation-level metaphors, we highlight previous work in this same direction in the next subsection.

Relation-level metaphor and similar phenomena 2.1

Apart from semantic and contextual features frequently exploited in state-of-the-art models, the construction grammar approach to metaphors (Sullivan 2013) has also highlighted the importance of syntax. Sullivan (2013) identified grammatical **constructions** (form-meaning or syntactico-semantic patterns) in which metaphors are frequently manifested. Identification approaches, most evidently those on the relation-level, most frequently concern metaphoric constructions of the following types: X is Y, ADJECTIVE-NOUN, NOUN-NOUN, SUBJECT-VERB, VERB-OBJECT.

Among earlier approaches, Kintsch (2000) proposes a computational model of "metaphor predication" for X is Y metaphors using LSA. They propose the "landmark method", in which they use cosine similarity to identify whether properties are transferred from source to target. Other works explore how static embeddings (e.g., word2vec, GloVe) can directly signal metaphor through semantic incongruity in constructions (Agres *et al.* 2016; Mao *et al.* 2018), and some extend the approach to the visual modality (Shutova *et al.* 2016). Shutova *et al.* (2010) present one of the first unsupervised approaches, identifying verbal metaphors in the BNC corpus by clustering noun and verb vectors from a seed set and extending source-target concepts through corpus-based vector similarity. Agres *et al.* (2016) evaluate word2vec and traditional count-based vectors on behavioral data to test if they encapsulate metaphoricity, familiarity, and meaningfulness. For both vector types, their results show that low values of metaphoricity were predictors of high cosine similarity. Su *et al.* (2017) combine cosine similarity with WordNet relations to identify nominal metaphors (X is Y, e.g., *Achilles is a lion*) in English and Chinese. They classify the relation as a metaphor if the similarity is lower than a predefined threshold and the concepts have no taxonomic relationship in WordNet. As the threshold values for the two languages are very different, this indicates language-specific baselines need to be determined. Another study also based on a threshold value is by Mao *et al.* (2018) who use CBOW and SkipGram embeddings as well as WordNet to identify metaphorical verbs. For each target verb, they find the best-fit synonym, hypernym, or hyponym in WordNet that matches the context of the sentence. Then, they compute the cosine similarity between the best-fit word and the target verb and classify it as metaphor if the similarity is lower than a threshold of 0.6, established on the basis of a development set.

Shutova *et al.* (2016) explore visual and linguistic embeddings for phrase-level metaphor prediction, finding that multimodal embeddings perform best, and that measuring similarity between individual words outperforms phrase embeddings in linguistic-only settings. Zayed *et al.* (2018) propose a semi-supervised approach for identifying metaphoric verbs in verb-noun phrases. They use a seed set of known verb-noun phrases where the verb is metaphoric. For a given candidate verb, i.e., an unlabelled example, they start by finding the most similar verbs in the seed set, and the nouns co-occurring with them in the seed set of phrases. They calculate the distance between the candidate noun in the unlabelled phrase to each of the nouns collected from the seed set. Finally, if the average of these distances is below a threshold value, they classify the candidate phrase as metaphoric. They experiment with two distance/similarity metrics and two static word embedding methods, achieving optimal results with GloVE and cosine distance.

Pedinotti *et al.* (2021) test the knowledge instilled in BERT models by applying the "landmark method" introduced in Kintsch 2000. They show that BERT encodes metaphor-relevant properties more clearly in lower layers, especially for conventional expressions, and note a performance drop in upper layers for creative metaphors. Brglez (2023) presents the only known approach to relation-level metaphor identification in Slovene. To determine the metaphoricity of adjective-noun and noun-noun constructions, Brglez (2023) uses cosine similarity of constituent words to classify the phrase as a metaphor if the similarity is below a threshold value. The study also compares static fastText and SloBERTa embeddings, which result in comparable performance, with lower layers of SloBERTa better suited to the task. However, the study is limited by a very small dataset of only 48 examples.

Somewhat similar to relation-level metaphors are multi-word expressions (MWEs) and idioms. Among automatic approaches to MWEs, Cordeiro *et al.* (2019) investigate English nominal compounds and their French and Portuguese counterparts. To distinguish compositional from non-compositional (idiomatic) MWEs, they measure the cosine similarity between the combined vectors of the parts and the vector of the compound. They find that the models can successfully capture idiomaticity, with word2vec as the best performing model for English, while for French and Portuguese, models based on association measures fared better. Garcia *et al.* (2021) investigate various contextual models for their representation of potentially idiomatic expressions, i.e., those that can be literal or idiomatic depending on the context, in English and Portuguese. They measure the cosine similarity of the embeddings of idiomatic compounds with 1) the embeddings of their meaning-preserving compounds and 2) literal synonyms of the components. They show that idiomatic phrases are closer to the literal synonyms than to their meaning-preserving paraphrases, concluding idiomaticity is not yet adequately captured by contextual models.

2.2 *Intrinsic analyses of language models*

Among the various approaches to analyzing the inner workings of language models, for example through attention mechanisms and probing (e.g., Voita *et al.* 2019; Liu *et al.* 2019; Hewitt and Manning 2019), here we report related work most relevant to our study, which is focused on the representation of semantic information.

Wiedemann *et al.* (2019) studied the representation of polysemous words in Flair, ELMo, and BERT. They find that contextualized embeddings place different senses of a word in different regions, especially in BERT. In the task of word sense disambiguation, both Loureiro *et al.* (2020) and Reif *et al.* (2019) find lower layers less effective for disambiguation than upper layers. This is in line with the study by Ethayarajh (2019) showing that embeddings of BERT, ELMo, and GPT-2 become increasingly more contextualized, i.e., context-specific in the upper layers. The author measures how similar a word's representation is to itself across various contexts, and reveal that as one moves from lower to upper layers, word representations shift from being more general and context-independent (high self-similarity across contexts) to more context-specific (low self-similarity in different contexts). The increasing contextualization effect was empirically validated by other studies. In the study of various pre-trained language models, Vulić *et al.* (2020) test and compare the performance of embeddings from different layers in various lexicosemantic tasks, such as word analogy or lexical relation prediction. Among other things, they recommend choosing monolingual LMs; encoding words with multiple contexts; and averaging over lower layers, as the latter seems to concentrate more type-level lexical information. A follow-up study by Burdick *et al.* (2022) re-evaluates the previous work by investigating the representation of words in paraphrases and studying the correlation of cosine similarity to human similarity judgments. The authors find that when controlling for the meaning of words, upper layers produce more similar representations, i.e., are more correlated with human similarity judgments of words in context. Namely, in BERT's lower layers, the cosine similarity between identical words is relatively high for the same word regardless of context. As one moves to higher layers, the similarity declines. While the decline for words that are the same in the two paraphrases and also positionally aligned (retain the meaning/function) is steady and gradual, the similarity decline for words which are not completely aligned in the two paraphrases (i.e. have different meanings/functions) happens earlier and is more pronounced. According to Wang and Zhang (2024) who deal with word embedding similarity in the context of WSD in different layers of contextual models, BERT-based models exhibit "first word position bias." In their experiments, the cosine similarity of two words that appeared at the start of the input sentences was considerably higher than the similarity of words that appeared in later positions. However, when simply prefixing and suffixing the input with quotation marks, the similarity dropped and led to higher accuracy. The effect of position has also been observed by Mickus *et al.* (2020) and Burdick *et al.* (2022).

The studies above all reflect how different layers capture varying degrees of semantic information. In line with our own approach, many of the studies have found lower layers of contextual models to be more like static embeddings in terms of stability across contexts and their usefulness for "type-level" tasks. They indicate that lower layers capture more surface-level features, stable across different contexts, much like the basic meaning of a word is stable out of context, and, conversely, that upper layers better match the final task objectives that require context information. Thus, we would expect to observe the most relevant semantic differences between the constituent words of metaphoric phrases in the lower layers of the model. However, as Vulić *et al.* point out, representations that work best are highly dependent on the task and language at hand. In this work, we elucidate the topic for Slovene by investigating both contextual and static representations for one specific type-level task: the representation of basic word meaning and basic meaning (dis)similarity in metaphoric and non-metaphoric constructions. Specifically, we address the following research questions:

1. Is it possible to identify relation-level metaphors via the incongruity of basic word meaning, specifically via the cosine similarity of word embeddings?

2. Are static and contextual embeddings comparable in terms of representing basic meaning incongruity in metaphoric phrases?
3. Are monolingual and multilingual models comparable in terms of representing basic meaning incongruity in metaphoric phrases?
4. Which layers of contextual embedding models better represent basic meaning incongruity in metaphoric phrases?
5. How is cosine similarity affected by syntax, i.e., is there any difference in constructions involving adjective-noun (amod) and noun-noun (nmod) relations?

3 METHODS

In this section, we present the methods used to investigate the basic meaning representations in static and contextual embeddings. First, we present the datasets on which we conducted our experiments and the decisions taken in order to select the sample data. Following is the description of the word embedding models used, the types of inputs to the models, the manner of retrieving the embeddings, and the metric used to calculate semantic similarity. The final part of our methodology concerns significance testing and a more detailed analysis of examples.

3.1 *Datasets*

Experiments in this study were carried out on two datasets. For the preliminary set of experiments on all models, we use the dataset previously presented in Brglez 2023, consisting of metaphoric and non-metaphoric pairs of phrases for 24 Slovene words (8 adjectives and 16 nouns). It includes three types of constructions: adjective-noun with a potentially metaphoric adjective; adjective-noun with a potentially metaphoric noun; and noun-noun, where the first noun can be metaphoric. Examples of the three types of phrases are shown in Table 1.[2] For example, the word *steber* 'pillar' is used literally in the phrase *sredinski steber* 'central pillar', and metaphorically in *moralni*

[2] The number of examples is equally distributed among construction types, i.e., 8 per type.

Construction	Example
NOUN_{met}-NOUN_{lit}	***oblaki*** *dvoma* '**clouds** of doubt'
NOUN_{lit}-NOUN_{lit}	***oblaki*** *metana* '**clouds** of methane'
ADJ_{met}-NOUN_{lit}	***prežvečena*** *fraza* '**chewed-up** phrase'
ADJ_{lit}-NOUN_{lit}	***prežvečena*** *hrana* '**chewed-up** food'
ADJ_{lit}-NOUN_{met}	*moralni* ***steber*** 'moral **pillar**'
ADJ_{lit}-NOUN_{lit}	*sredinski* ***steber*** 'central **pillar**'

Table 1:
Examples from the Brglez 2023 dataset: lit = literal use, met = metaphoric use

steber 'moral pillar'. For each of these phrases, the dataset also contains one sentence sampled from Gigafida (Krek *et al.* 2019), a reference corpus of Slovene.

For further experimentation, we collect examples from the Slovene metaphor corpus KOMET 1.0 (Antloga 2020a,b). The whole corpus contains about 200,000 words in 14,000 sentences and is similar in size and genre makeup to the English metaphor corpus VUAMC (Steen 2010). The corpus was automatically linguistically annotated for lemmas, part-of-speech, syntactic dependencies, and morphosyntactic tags, and manually annotated for metaphors according to a modified version of the MIPVU guidelines (Steen 2010) by one person.[3] The manual annotations differentiate among three types of metaphor-related words (`MRWs`): indirect, direct, and implicit, as well as metaphor flags (signals of metaphoricity), idioms, metonymies, and adverbial phrases. Indirect metaphors are lexical units that have a contextual sense that differs from their most basic sense. That is, the referent in the context is different from the referent this word would usually have. In (1) below, the words *v* 'in', *izgubil* 'lost', *na* 'on', *od* 'from', *mladih* 'young', *nog* 'feet' are marked as indirect metaphors (`MRWi`).[4] For example, the prepositions *v* 'in', *na* 'on', *od* 'from' would usually refer to an object or location in space, however, they express more

[3] The modifications stem from the lack of a corpus-based dictionary and the lack of certain linguistic phenomena, such as phrasal verbs. The procedure for Slovene has not yet been formalized. Since our analysis revealed numerous erroneous metaphor annotations, a random sample of 4000 tokens was additionally annotated by an expert linguist (see Section 5.2).

[4] In the examples cited from KOMET, the relevant nmod or amod constructions are in bold, and other relevant parts of the sentence are underlined. For brevity and readability, we omit or abbreviate some annotations.

abstract relations in this context. Another example is the verb *izgubil* ‘lost’, which would usually mean ‘to no longer have something or know where it is’; however, in this context it stands for ‘to have a person taken away by death’. Such indirect metaphors also have additional annotations in KOMET, consisting of the semantic/lexical field that acts as a source domain in a given metaphor.[5]

(1)

V	prometni	nesreči	sem	izgubil	brata,	na	katerega
in	traffic	accident	am-AUX	lost	brother,	on	whom
`MRWi`				`MRWi`		`MRWi`	

sem	bil	<u>od</u>	<u>**mladih**</u>	<u>nog</u>	zelo	navezan.
am-AUX	was	from	young	feet	very	attached.
		`MRWi`	`MRWi`	`MRWi`		
		`<`	`#met.idiom`	`>`		

‘I lost my brother, to whom I was very attached from a young age, in a traffic accident.’

Direct metaphors are lexical units whose contextual and basic senses are the same but which are incongruous with the topic domain, and where some sort of cross-domain mapping is detectable. They include comparisons and similes, and may be signaled via “metaphor flags” such as ‘like’ or ‘literally’. Such a case can be observed in (2) below, where a loud motor is being directly compared via the metaphor flag (`MFlag`) *kot* ‘like’ to *stara barkača* ‘old trawler’. The latter still refers to and evokes the image and characteristics of an old fishing boat, but the utterance creates a direct cross-domain mapping from the boat to the motor.

(2)

V	nizkih	obratih	je	njegov	motor	ropotal	kot
in	low	revolutions	is-AUX	his	motor	rumbled	like
`MRWi`							`MFlag`

stara	**barkača.**
old	trawler.
`MRWd`	`MRWd`

‘At lower revs, his motor rumbled like an old trawler’.

[5] These additional tags include, for example, `#met.personification`, `#met.purposive_area`, `#met.spatial_orientation`, `#met.motion`.

A third class of metaphors annotated in the corpus are implicit metaphors, which encompass words such as pronouns that stand for metaphorically used words. Words marked as implicit metaphors are thus not metaphoric themselves but stand in for other metaphor-related words and function as cohesive devices. An implicit metaphor from the corpus can be observed in (3), where the word *jim* 'them' is annotated as an implicit metaphor (`MRWimp`) because it refers to the previously metaphorically used phrase *koščki mozaika* 'pieces of mosaic'.

(3)									
	/.../	ti	**koščki**	**mozaika**,	ki	jim	rečem	družina	/.../
	/.../	these	pieces	mosaic,	which	them	say.1SG	family	/.../
			`MRWd`	`MRWd`		`MRWimp`	`MFlag`		

'/.../ these pieces of mosaic which I call family /.../'

For the purposes of this experiment, we extract only sentences and constructions that follow the same relation-level paradigm, namely noun phrases with an adjectival or nominal modifier (ADJ-NOUN and NOUN-NOUN constructions). The extraction procedure leverages the existing dependency annotations and extracts all sentences that feature *amod* (adjective modifier) and *nmod* (noun modifier) syntactic relations. This amounts to 9,519 sentences. In these, we find 34,264 such word pairs, out of which 29,844 are unique. In Table 2, we can see that a total of 31,934 of those pairs are completely literal, that is, none of the two words in the relation is labelled as metaphoric, and in 2,330 cases, at least one of the words has a metaphor-related tag.

Use	amod	nmod	Total
literal	19,472	12,462	31,934
metaphoric	1,269	1,061	2,330
Total	20,741	13,523	**34,264**

Table 2: Construction counts from KOMET 1.0

As a first step, we exclude constructions containing proper nouns such as *Anglež Whitwell* 'the Englishman Whitwell' because we presume these are not well represented in the vector space and could produce unwanted noise both for literal and metaphoric pairs. Moreover, there are issues with a simple delimitation of the data relying only on whether a relation contains a metaphor-related-word (`MRW`) tag or a metaphor frame tag. The issues we address are:

1. The corpus includes idioms and fixed adverbial expressions as metaphor-related words, which may or may not be metaphoric, and may only partially correspond to the extracted noun phrase (e.g., only one of the words is part of the idiom). In (1), the extraction samples *mladih nog* ‘young feet’ as a metaphor-containing construction. However, these are only tagged because they are part of a wider idiomatic expression *od mladih nog* ‘from a young age’. On its own, it could be argued that the phrase only exhibits metonymy, as it associates the age of the person with the age of their limbs.
2. We observe that words referring to non-human entities that are personified by other words are marked as metaphor-related words through the frame tag `#met.personification`. In (4) below, only the noun *stezo* ‘path’ is marked as a `MRW`, however, the actual metaphor in the sentence is the verb *pripeljala* ‘lead, drive’ that personifies the noun by giving it human-like abilities, i.e., to lead, drive someone somewhere.

(4)

Kmalu	so	zapeljali	na	še	**ožjo**	**stezo**,	ki	pa
soon	are-AUX	drove	on	even	narrower	path,	which	but
						`MRWi`		
						`#met.person.`		

jih	je	hitro	pripeljala	do	čudovite	gozdne	jase.
them	is-AUX	quickly	led	to	beautiful	forest	clearing.

‘Soon, they drove onto an even narrower path that quickly led them to a beautiful forest clearing.’

3. In instances where both constituent words are marked as metaphors, the construction is only metaphoric if embedded in a wider incongruous context, and is not metaphoric on the relation between the two constituents. Thus, for the purposes of this experiment, such constructions are considered **literal**. Sometimes overlapping with this phenomenon are direct metaphors, such as

stara barkača ‘old trawler’ from (2) or *koščki mozaika* ‘pieces of mosaic’ from (3).

4. Another issue is phrases marked as adverbial phrases, i.e., multi-word expressions that perform the function of an adverb. In (5) below, the adverbial phrase is composed of a preposition *po* and noun *zaslugi*, the latter also being part of an extracted nmod relation.

(5)

Vsekakor	gre	za	izjemno	odkritje,	do	katerega
absolutely	goes	for	extraordinary	discovery,	to	which
		MRWi			MRWi	

je	prišlo	predvsem	po	**zaslugi**	**amaterjev**,	ki
is-AUX	came	especially	after	merit	amateurs,	who
			MRWi <#met.adverbial_ph.>			

so	prečesavali	javno	dostopne	podatke.
are-AUX.3PL	combed.through	publicly	available	data.
	MRWi			

‘Certainly it is an exceptional discovery arrived at especially thanks to non-professionals who combed through publicly available data.’

5. As previously mentioned, tags in the KOMET corpus also include metonymies. These differ from metaphors in a crucial aspect: while metaphors equate or compare two unrelated concepts or domains on the basis of similarity, metonymies are expressions that substitute concepts with others from the same domain on the basis of association or inclusion. By using metonymy, we use a concept that is associated with, includes, or is included in another concept.

We thus decide to filter out such instances, resulting in 628 nmod metaphor-containing relations and 813 amod metaphor-containing relations. Our final test sample, however, contains 591 amod metaphors and 738 nmod metaphors, as a total of 112 metaphoric examples contained out-of-vocabulary words for our most limited static embedding model. In addition to the metaphoric examples, we randomly sample the same number of literal phrases (those without metaphor-related words),[6] which amounts to a total of 2658 examples or phrase-sentence pairs (Table 3).

Table 3:
Number of constructions sampled from KOMET 1.0

Use	amod	nmod	Total
literal	591	738	1329
metaphoric	591	738	1329
Total	1182	1476	2658

3.2 *Word embedding models*

We compare word embeddings obtained by two main methods: static and dynamic. For static embeddings, we use 100-dimensional CLARIN.SI-embed.sl fastText embeddings (fT_CLARIN, Ljubešić and Erjavec 2018), 300-dimensional EMBEDDIA fastText embeddings (fT_EMBEDDIA),[7] and 1024-dimensional word2vec-like embeddings for 200,000 words obtained from their average ELMo representations (w2v_ELMo).[8] For contextual embeddings, we include encoder-only, encoder-decoder, and decoder-only models including ELMo and different Transformer-based models. We experiment with both monolingual and multilingual models.

Among monolingual models, we include SloBERTa 2.0 (Ulčar and Robnik-Šikonja 2021), a Slovene pre-trained RoBERTA model. Among the contextualized embedding models for Slovene, this architecture has performed best in most monolingual tasks (Ulčar *et al.* 2021). We also include Slovene ELMo (Ulčar and Robnik-Šikonja 2020) and two

[6] The random sampling of literal phrases was done independently of the individual metaphoric phrases. Namely, the words do not necessarily appear in the KOMET corpus in both literal and metaphorical use, which is why we could not ensure that the samples consist of metaphorical-literal pairs as in the dataset by Brglez (2023).

[7] The embeddings are available upon request from the EMBEDDIA project collaborators.

[8] The first LSTM layer is used to produce the vector values in word2vec format. The embeddings are available upon request from Andraž Repar, Aikwit.

Table 4: Overview of the models used in the study

	Embedding	Language support	Layers	Dimensions
Static	fT_CLARIN	1		100
	fT_EMBEDDIA	1		300
	w2v_ELMo	1		1024
Dynamic	ELMo	1	3	1024
	SloBERTa	1	12	768
	CroSloEngual BERT	3	12	768
	XLM-r-slobertić	5	24	1024
	sloT5-small	1	8	512
	sloT5-large	1	24	1024
	mT5-small	101	8	512
	mT5-large	101	24	1024
	text-embedding-ada-002	?	?	1536
	text-embedding-3-small	?	?	1536
	text-embedding-3-large	?	?	3072

Slovene versions of the T5 encoder-decoder, sloT5-small and sloT5-large (Ulčar and Robnik-Šikonja 2023). Multilingual models can be separated into those trained on a smaller set of languages including CroSloEngual BERT (CSE BERT, Ulčar and Robnik-Šikonja 2020) trained on Croatian, Slovene, English, and XLM-r-slobertić (Ljubešić *et al.* 2024) trained on Croatian, Bosnian, Montenegrin, Serbian, Slovene,[9] and massively multilingual models including mT5-small and mT5-large (Xue *et al.* 2021), as well as GPT-based embeddings text-embedding-ada-002 (OpenAI 2022), text-embedding-3-small, and text-embedding-3-large (OpenAI 2024). The main characteristics of the models are laid out in Table 4.[10]

3.3 *Input methods and preprocessing*

In the static embedding setting, we obtain only one embedding per word, as they are context-insensitive. To obtain word embeddings

[9] The model was created by additional pre-training of a multilingual XLM-ROBERTA model, originally trained on 100 different languages.

[10] OpenAI does not provide detailed implementations of their embedding models hence the ? character in the columns Language support and Layers.

from contextual models, we test different context settings. Following Brglez (2023), we test whether the models, like humans, embed and represent the most basic meaning of a word if the latter is presented individually, without any context. Conversely, we would expect that a word embedded in the context of a metaphorical construction or a full sentence will contain a more contextual, shifted meaning of the word. Thus, as reported in Brglez 2023, the input to the model is one of:

1. `no context` (**I**). The input to the model is just the individual word (two separate inputs per phrase).
2. `phrase context` (**P**). The input to the model is the whole phrase.
3. `sentence context` (**S**). The model is presented the complete sentence that exemplifies the use of a phrase.

The experiments in Brglez 2023 were limited by the small and controlled set of examples. Conversely, we test our hypotheses on a larger dataset that contains constructions that are structurally more diverse. Namely, while the example constructions from Brglez 2023 only contain two words, the larger dataset includes constructions where the modifier and head of a particular grammatical relation are many words apart and can thus greatly vary in length. Thus, in addition to the previously mentioned input types, we also introduce the word pair (WP) input (which may sometimes overlap with the phrase (P) input).

4. `word pair` (**WP**). The input to the model is only the two words, not necessarily contiguous, from the relation: the modifier and the head. Such pairs are less collocational than contiguous pairs and will allow us to compare the performance of the models on phrasal (P) versus non-phrasal (WP) metaphors.

Additionally, because of the possible "first word position bias" (Wang and Zhang 2024), we prepend each of the inputs with a simple prefix *Primer:* 'Example:'. To avoid issues arising from the different tokenization strategies by various models employed, we demarcate punctuation characters with whitespace to prevent models from including punctuation in tokens. Below is a demonstration of the different inputs to the model for the phrase *pekel ameriškega Divjega zahoda* 'hell of the American Wild West':

1. **I**: *Primer : pekel*; *Primer : zahoda*
2. **WP**: *Primer : pekel zahoda*
3. **P**: *Primer : pekel ameriškega Divjega zahoda*
4. **S**: *Primer : O njej in drugih podobnih pustolovščinah so pisali že Salinger , Kerouac in Hunter S. Thompson , a nihče od njih nam ni upal priznati, da ameriški sen med zaspanimi mesti , kanjoni in divjo puščavo , ki je nekoč predstavljala pekel ameriškega Divjega zahoda , nikoli ni zares obstajal .*

   'Example : Salinger, Kerouac, and Hunter S. Thompson have previously written about this adventure or a similar one , but none of them dared to admit to us that amidst the sleepy towns , canyons and the wild desert , which used to represent the hell of the American Wild West , the American dream never really existed .'

The only exception for the input types, the pre-pending strategy, and the averaging of subword tokens are the GPT-based text-embedding-ada-002, text-embedding-3-small, and text-embedding-3-large. These text embedding models only generate one embedding regardless of the number of tokens, be it a word or a whole text paragraph. To allow us to compute the cosine similarity between words, here we only use the I input strategy to obtain the embeddings for individual words.

Embedding retrieval 3.4

We experiment with embeddings obtained separately from each layer, namely the embedding layer, the input layer, and all subsequent hidden layers. For ELMo, embeddings for layers 0, 1, and 2 are obtained by full-weighing the relevant layer and zero-weighing the non-relevant layers using the AllenNLP Library.[11] For open transformer models, we use the huggingface transformers library to access the models and obtain embeddings from hidden states from each of the layers.[12] For the closed GPT system, we only obtain the embeddings returned by the OpenAI API, which are presumably the embeddings from the final layer.[13] The vector of words that are split into subword

[11] https://allenai.org/allennlp/software/allennlp-library
[12] https://huggingface.co/
[13] https://platform.openai.com/docs/api-reference/introduction

tokens during tokenization is obtained from the element-wise mean of all its subword tokens.

The initial token embeddings, which serve as the initial input to the model, can be considered static, as they are not yet contextualized – that is, they do not change based on surrounding words. In models like ELMo and T5, the embeddings at layer 0 are identical to these initial token embeddings and remain the same regardless of context. Thus, we report only one result for these models at the input layer. However, in BERT-based models like XLM-R and SloBERTa, the input to the 0th layer also includes positional encoding, which alters the initial token embedding by adding information about the token's position. This is why we report one result for ELMo and T5 on the input layer, and separate results for each input to the embedding and input layers of other models.

3.5 *Similarity metric*

Words participating in metaphoric constructions originate in different conceptual/semantic domains. We therefore expect them to exhibit less semantic similarity and thus be further apart in the vector space of embeddings than the constituents of non-metaphoric constructions. To measure semantic similarity, we apply the frequently used cosine similarity metric (see (6)), which estimates the similarity of two words through the cosine of the angle between the embeddings of the words, where $\mathbf{v_1}$ and $\mathbf{v_2}$ are the word embedding vectors of two words:

(6) $$\text{cos\_sim}(\mathbf{v_1}, \mathbf{v_2}) = \frac{\mathbf{v_1} \cdot \mathbf{v_2}}{||\mathbf{v_1}||||\mathbf{v_2}||}$$

3.6 *Significance testing*

We test the significance of cosine similarities by comparing two distributions: one containing similarities between words in metaphoric constructions, and the other containing similarities from non-metaphoric constructions. Each construction in the dataset is treated as an independent sample. Significance testing is performed separately for all pairs of distributions in all the different input and layer combinations described in the previous sections. For the first experiment, we only use the smaller dataset of 48 word pairs and the input types as in Brglez 2023 and calculate the significance of the difference in means of the two distributions (metaphoric and non-metaphoric cosine similarities) using Student's T test (Student 1908). On the larger sample collected from the KOMET corpus, we also use the independent t-test (Student 1908), where we separately compare metaphoric and non-metaphoric adjective-modifier-constructions, as well as metaphoric and non-metaphoric noun-modifier constructions in various contexts. We test each set of data for variance, and in case of unequal variance, we employ the modified Welch's t-test (Welch 1947). The reported effect size is Cohen's d (Cohen 2013), calculated with respect to the t-test variant employed. The effect size d can range from 0 to 2, where 0 is complete overlap and 2 is complete divergence. The values in between can be interpreted more finely as d (.01) = very small, d (.2) = small, d (.5) = medium, d (.8) = large, d (1.2) = very large, and d (2.0) = huge effect (Sawilowsky 2009).

Analysis of tail-end examples 3.7

Metaphors can range from novel to conventional, from creative ones coined on-the-fly to those frequently used and thus earning their lexicalized place in dictionaries. In the latter case, they can pass unnoticed as they may appear as "literal" phrasings, part of the ordinary vocabulary. As language model representations rest upon the distributions of words in the texts they are trained on, we might also find the frequency of co-occurrence, connected to the concept of novelty/conventionality of metaphors, to have an effect on cosine similarity.

To test whether this is the case for our distributions, one part of the analysis concerns the effect of collocability on cosine similarity. To this end, we investigate whether the phrases sampled from the distribution tails appear in the frequency list of collocations (Krek *et al.* 2021).[14] From this resource, we can indirectly derive whether we are perhaps dealing with more novel or more conventional examples. Because of the high morphological richness of Slovene, we match

[14] The resource contains collocations in 81 predefined syntactic structures which appear in the Slovene reference corpus Gigafida 2.1 at least 10 times.

sampled constructions and collocations at the level of lemmas. We consider perfect and fuzzy matches, meaning we consider examples with words inserted between the two lemmas of the relation (e.g., the construction *pogled v prihodnost* 'look into (the) future' matches two collocations, *pogled in prihodnost* 'look and (the) future', *pogled v prihodnost* 'look into (the) future'), which we manually validate. In the second part of the analysis, we manually study and annotate the examples to gain insights into why some pairs appear at the left and others at the right tail of the cosine similarity distribution.

4 RESULTS AND DISCUSSION

4.1 *Cosine similarity and t-testing on toy dataset*

In the first experiment, we perform significance testing for the cosine similarities of 48 metaphoric vs. non-metaphoric noun phrases. As mentioned above, this dataset was constructed manually and is a collection of "ideal" examples, in which the distinction between metaphoric and literal use can be very easily determined. Tables 5 and 6 show the t-scores and the associated p values for static and GPT embeddings, while Table 7 shows the results for cosine similarities of embeddings obtained from each of the layers.

Table 5:
T-test results on the Brglez 2023 dataset for static embeddings. Results with $p < .001$ in bold italic

Embedding	t (p)
fT_EMBEDDIA	***3.52 (<.001)***
fT_CLARIN	***3.63 (<.001)***
w2v_ELMo	***4.47 (<.001)***

Table 6:
T-test results on the Brglez 2023 dataset for GPT-based embeddings

Model	t (p)
text-embedding-3-small	0.68 (.50)
text-embedding-3-large	2.24 (.03)
text-embedding-ada-002	0.58 (.57)

Table 7: T-test results on the Brglez 2023 dataset for contextual embeddings. I = individual word inputs, P = complete phrase input, S = sentence input. Results with $p < .01$ in bold, $p < .001$ in bold italic

Model	Layer / Input type	0	1	2	3	4	5	6	7	8	9	10	11	12
	I	**2.73 (<.01)**	***3.72 (<.001)***	***3.68 (<.001)***	***3.58 (<.001)***									
ELMo	P	**2.73 (<.01)**	***4.31 (<.001)***	***3.56 (<.001)***	***3.84 (<.001)***									
	S	**2.73 (<.01)**	**3.19 (<.01)**	***3.91 (<.001)***	***3.93 (<.001)***									
	emb	1.33 (.19)												
	I	**3.00 (<.01)**	***3.34 (<.01)***	*2.26 (.03)*	1.75 (.09)	*2.30 (.03)*	2.28 (.03)	2.11 (.04)	2.35 (.02)	2.48 (.02)	2.63 (.01)	2.30 (.03)	2.33 (.02)	0.57 (.57)
SloBERTa	P	**2.78 (<.01)**	**2.98 (<.01)**	**2.77 (<.01)**	**2.47 (<.01)**	**3.16 (<.01)**	**2.98 (<.01)**	**3.01 (<.01)**	**2.91 (<.01)**	2.55 (.01)	2.23 (.03)	1.58 (.12)	*1.36 (.18)*	0.84 (.41)
	S	**2.85 (<.01)**	***3.73 (<.001)***	***3.55 (<.001)***	2.58 (.01)	2.30 (.03)	2.17 (.03)	**2.84 (<.01)**	**2.69 (<.01)**	2.45 (.02)	1.94 (.06)	1.71 (.09)	0.62 (.54)	1.47 (.15)
	emb	0.27 (.79)												
	I	0.44 (.66)	0.94 (.35)	1.51 (.14)	1.89 (.06)	2.44 (.02)	2.61 (.01)	1.84 (.07)	1.78 (.08)	1.45 (.15)	1.17 (.25)	1.21 (.23)	1.30 (.20)	0.90 (.37)
CSE BERT	P	0.53 (.60)	1.09 (.28)	1.85 (.07)	2.25 (.03)	2.46 (.02)	2.01 (.05)	1.48 (.14)	1.56 (.13)	1.13 (.26)	0.98 (.33)	1.21 (.23)	1.07 (.29)	0.76 (.45)
	S	0.55 (.59)	1.35 (.18)	2.63 (.01)	2.39 (.02)	2.39 (.02)	1.64 (.11)	0.55 (.58)	0.27 (.79)	0.50 (.62)	0.64 (.52)	0.28 (.78)	0.22 (.83)	0.47 (.64)
	emb	1.08 (.28)												
	I	1.70(.10)	0.89 (.38)	1.07 (.29)	0.44 (.66)	0.24 (.81)	0.60 (.55)	0.40 (.69)	0.54 (.59)	0.31 (.76)	0.46 (.65)	0.29 (.77)	0.07 (.95)	0.34 (.74)
xlm-r-slobertić	P	1.92 (.06)	0.88 (.38)	0.99 (.33)	0.33 (.74)	0.52 (.60)	1.02 (.31)	1.53 (.13)	1.01 (.32)	0.47 (.64)	0.01 (.99)	0.12 (.90)	0.06 (.95)	0.96 (.34)
	S	2.44 (.02)	0.91 (.37)	1.14 (.26)	0.35 (.73)	0.27 (.79)	0.61 (.54)	1.02 (.31)	1.21 (.23)	1.14 (.26)	2.52 (.02)	2.48 (.02)	2.59 (.01)	1.97 (.05)
	(cont.)		**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
I			-1.11 (.27)	-1.96 (.06)	-2.09 (.04)	-2.13 (.04)	-2.20 (.03)	-2.56 (.01)	-1.37 (.18)	0.22 (.82)	0.68 (.50)	0.48 (.63)	0.46 (.65)	-0.47 (.64)
P			0.68 (.50)	0.53 (.60)	0.52 (.61)	0.35 (.73)	0.4 (.69)	0.35 (.73)	-0.02 (.98)	0.19 (.85)	-0.84 (.41)	-0.79 (.43)	-0.73 (.47)	-0.28 (.78)
S			-1.7 (.10)	-1.10 (.28)	-1.13 (.26)	-0.15 (.88)	0.04 (.97)	0.12 (.91)	-0.52 (.61)	-0.34 (.74)	0.63 (.53)	-0.41 (.68)	-0.75 (.46)	-1.59 (.12)
	I	***3.76 (<.001)***	1.69 (.10)	1.49 (.14)	1.95 (.06)	1.66 (.10)	0.92 (.36)	0.70 (.49)	0.49 (.63)	1.17 (.25)				
sloT5-small	P	***3.76 (<.001)***	**2.88 (<.01)**	***3.55 (<.001)***	**3.03 (<.01)**	**2.95 (<.01)**	**3.14 (<.01)**	2.28 (.03)	1.82 (.07)	1.08 (.28)				
	S	***3.76 (<.001)***	**3.18 (<.01)**	**2.87 (<.01)**	**2.69 (<.01)**	***3.70 (<.001)***	**3.02 (<.01)**	**2.94 (<.01)**	2.15 (.04)	**3.44 (<.01)**				
	I	**3.46 (<.01)**	***4.02 (<.001)***	***4.39 (<.001)***	***4.50 (<.001)***	***4.71 (<.001)***	***4.90 (<.001)***	***4.80 (<.001)***	***4.59 (<.001)***	***4.58 (<.001)***	***4.61 (<.001)***	***4.51 (<.001)***	***4.44 (<.001)***	***4.02 (<.001)***
	P	**3.46 (<.01)**	***4.80 (<.001)***	***5.35 (<.001)***	***5.40 (<.001)***	***5.56 (<.001)***	***5.77 (<.001)***	***5.76 (<.001)***	***5.68 (<.001)***	***5.58 (<.001)***	***5.58 (<.001)***	***5.45 (<.001)***	***5.55 (<.001)***	***4.96 (<.001)***
	S	**3.46 (<.01)**	***4.41 (<.001)***	***4.23 (<.001)***	***4.82 (<.001)***	***4.87 (<.001)***	***4.79 (<.001)***	***4.73 (<.001)***	***4.73 (<.001)***	***4.68 (<.001)***	***4.85 (<.001)***	***5.04 (<.001)***	***5.26 (<.001)***	***5.10 (<.001)***
sloT5-large	(cont.)		**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
	I		***3.88 (<.001)***	***3.66 (<.001)***	**3.36 (<.01)**	***3.56 (<.001)***	**3.29 (<.01)**	**3.27 (<.01)**	**3.03 (<.01)**	**3.01 (<.01)**	**2.71 (<.01)**	2.50 (.02)	2.30 (.03)	**3.07 (<.01)**
	P		***5.18 (<.001)***	***4.38 (<.001)***	***3.92 (<.001)***	***4.46 (<.001)***	***4.25 (<.001)***	***4.35 (<.001)***	***4.45 (<.001)***	***4.15 (<.001)***	***3.84 (<.001)***	**3.06 (<.01)**	2.51 (.02)	***3.53 (<.001)***
	S		***5.29 (<.001)***	***4.87 (<.001)***	***4.74 (<.001)***	***5.30 (<.001)***	***5.18 (<.001)***	***5.13 (<.001)***	***4.63 (<.001)***	***4.81 (<.001)***	***4.79 (<.001)***	***4.64 (<.001)***	***4.29 (<.001)***	***4.38 (<.001)***
	I	0.24 (.81)	0.14 (.89)	0.17 (.87)	0.30 (.76)	0.33 (.75)	0.37 (.71)	0.40 (.69)	0.39 (.70)	0.49 (.63)				
mT5-small	P	0.24 (.81)	0.02 (.99)	0.17 (.86)	0.15 (.89)	0.28 (.78)	0.37 (.72)	0.56 (.58)	0.74 (.47)	0.74 (.46)				
	S	0.24 (.81)	0.74 (.46)	0.49 (.62)	0.77 (.44)	0.63 (.53)	0.74 (.46)	0.25 (.80)	0.21 (.83)	0.47 (.64)				
	I	0.42 (.68)	0.11 (.91)	0.38 (.70)	0.80 (.43)	0.72 (.48)	0.40 (.69)	0.41 (.69)	0.30 (.77)	0.19 (.85)	0.40 (.69)	0.68 (.50)	0.60 (.55)	0.34 (.74)
	P	0.42 (.68)	0.13 (.90)	0.22 (.83)	0.74 (.46)	0.59 (.56)	0.44 (.66)	0.65 (.52)	0.52 (.60)	0.36 (.72)	0.66 (.51)	1.04 (.30)	1.39 (.17)	0.95 (.35)
	S	0.42 (.68)	0.01 (1.00)	0.36 (.72)	0.42 (.67)	0.39 (.70)	0.44 (.66)	0.63 (.53)	0.70 (.48)	0.65 (.52)	0.59 (.56)	1.05 (.30)	1.17 (.25)	1.45 (.15)
mT5-large	(cont.)		**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
	I		0.14 (.89)	0.16 (.87)	0.06 (.95)	0.19 (.85)	0.24 (.82)	0.01 (.99)	1.04 (.30)	1.73 (.09)	1.69 (.10)	1.01 (.32)	1.00 (.32)	0.36 (.72)
	P		0.89 (.38)	0.91 (.37)	0.38 (.71)	0.49 (.63)	0.04 (.97)	0.20 (.84)	0.21 (.84)	0.10 (.92)	0.38 (.71)	1.07 (.29)	1.12 (.27)	0.74 (.46)
	S		0.97 (.34)	0.96 (.34)	0.61 (.55)	0.50 (.62)	0.60 (.55)	0.34 (.73)	0.30 (.76)	0.08 (.94)	0.21 (.83)	0.45 (.66)	0.18 (.86)	0.25 (.81)

We can observe that while metaphoric vs. non-metaphoric pairs are statistically different in all the static models (Table 5), with $p < 0.001$ in all the cases, embeddings obtained from contextual models (Tables 6 and 7) exhibit somewhat less of a difference. In particular, there seems to be no statistically significant difference in most embeddings of multilingual models, including massively multilingual (mT5, GPT embeddings) and those trained on a small set of languages (xlm-r-slobertić, CroSloEngual BERT). On the other hand, the results on all monolingual models suggest it is possible to discern a statistical difference between the distribution of cosine similarities in metaphoric and non-metaphoric pairs. For ELMo, this is true on all layers for all input types, with the largest difference on layer 1, phrase input ($t = 4.31$ $p < .001$). For SloBERTa, it is most evident at the starting layers in embeddings created from a phrase input. The largest difference is shown on layer 1 in embeddings created from sentences ($t = 3.73$, $p < .001$). The smaller sloT5 model seems to create a good semantic embedding on the input (layer 0), but the differences fade in the upper layers, especially in the case of individual word input (I). The most persistent differences are leveraged from embeddings from sloT5-large, where almost all the cosine similarities are different with statistical significance. The largest t-values are achieved in this model, the top ones on the phrase input in layers 2–11 and 13, and on the sentence input in layers 10–13 and 16–18.

4.2 *Cosine similarity in KOMET*

For further experimentation, we keep the models with statistically significant results (determined at $p < .01$ in t-testing) from the previous section and test them on the larger set of examples from KOMET. This includes all the static embedding models and Slovene ELMo, SloBERTa, sloT5-small, sloT5-large. In a preliminary experiment, we find adjective-modifier (amod) and noun-modifier (nmod) constructions exhibit different magnitudes of cosine similarity due to their contrasting syntactic relations and selectional constraints, which is why in this section, we study them separately.

In Table 8, results of the t-test are presented for static embeddings. We see that the difference in cosine similarity in metaphoric

Embedding	Relation	t *(p)*
fT_EMBEDDIA	amod	***7.43 (<.001)***
	nmod	***7.43 (<.001)***
fT_CLARIN	amod	***6.89 (<.001)***
	nmod	***5.86 (<.001)***
w2v_ELMo	amod	2.49 (0.01)
	nmod	1.82 (.06)

Table 8:
T-test results on the KOMET 1.0 dataset for static embeddings. Results with $p < .001$ in bold italic

versus non-metaphoric constructions in KOMET is statistically significant except for the w2v_ELMo embeddings. The largest difference in the two distributions is observed in fT_EMBEDDIA embeddings, where $t = 7.43$ ($p < .001$) for both adjective-modifier and noun-modifier constructions.

Similarly, the results show statistical significance in many experimental setups for contextual embeddings, separately for amod (Table 9) and nmod (Table 10) constructions. For ELMo, embeddings from all the layers show statistical significance. For both adjective modifier and noun modifier constructions, the most divergent seem to be the embeddings on the input layer with $t = 7.33$ ($p < .001$) and $t = 6.81$ ($p < .001$), respectively. SloBERTa embeddings show statistically significant differences especially in the lower layers (0–2). For nmod constructions presented in a sentence input, discernible differences persist throughout the layers except for the final one. The greatest difference is observed in layer 2, S input, for amod constructions ($t = 5.10$, $p < .001$), and on layer 1, S input for nmod constructions ($t = 6.43$, $p < .001$). Interestingly, we can observe a statistically significant negative t-test value in upper layers (10–12) of SloBERTa for adjective modifier constructions, meaning the distributions are reversed and metaphoric pairs are thus closer in terms of cosine similarity than literal pairs. In the case of sloT5-small, the differences in cosine similarities are significant in almost all layers with the I and S input, whereas for the WP and P inputs, the distributions of cosine similarities converge in the upper layers. The statistically most different embeddings for both types of constructions are generated by layer 1, S input ($t = 7.34$, $p < .001$; $t = 7.32$, $p < .001$). In the largest model, sloT5-large, cosine similarities from practically all combinations of input

Table 9: T-test results comparing cosine similarities of metaphoric and non-metaphoric adjective-modifier constructions from KOMET 1.0. I = individual word inputs, WP = word pair input, P = complete phrase input, S = sentence input. Results with $p<.01$ in bold, $p<.001$ in bold italic. The highest difference in distributions per model is underlined

Model	Layer / Input	0	1	2	3	4	5	6	7	8	9	10	11	12
ELMo	I	***7.33 (<.001)***	***5.87 (<.001)***	***5.21 (<.001)***	***6.26 (<.001)***									
	WP		***5.95 (<.001)***	***4.39 (<.001)***	***5.56 (<.001)***									
	P		***5.67 (<.001)***	***4.39 (<.001)***	***5.41 (<.001)***									
	S		***4.01 (<.001)***	***5.61 (<.001)***	***5.75 (<.001)***									
SloBERTa	emb	***3.91 (<.001)***												
	I	***5.02 (<.001)***	***3.40 (<.001)***	2.51 (0.01)	-0.82 (0.41)	-1.58 (0.11)	-2.20 (0.03)	-1.81 (0.07)	-1.87 (0.06)	-1.32 (0.19)	-1.76 (0.08)	-2.14 (0.03)	-2.44 (0.01)	-2.29 (0.02)
	WP	***4.76 (<.001)***	***4.93 (<.001)***	***4.20 (<.001)***	-0.51 (0.61)	-1.61 (0.11)	-1.63 (0.10)	-1.67 (0.09)	-1.13 (0.26)	-0.84 (0.40)	-0.50 (0.62)	**-3.21 (<.01)**	***-4.25 (<.001)***	***-3.49 (<.001)***
	P	***4.75 (<.001)***	***4.83 (<.001)***	***4.55 (<.001)***	0.13 (0.89)	-0.87 (0.39)	-0.80 (0.42)	-0.60 (0.55)	-0.44 (0.66)	-0.21 (0.83)	0.60 (0.55)	-1.61 (0.11)	**-2.61 (<.01)**	-2.25 (0.02)
	S	***4.63 (<.001)***	***4.81 (<.001)***	***5.10 (<.001)***	1.94 (0.05)	0.95 (0.34)	1.35 (0.18)	1.15 (0.25)	1.20 (0.23)	0.89 (0.37)	0.09 (0.93)	-0.45 (0.66)	-1.42 (0.15)	-1.04 (0.30)
sloT5-small	I	***4.43 (<.001)***	***5.29 (<.001)***	***6.53 (<.001)***	***5.27 (<.001)***	***5.15 (<.001)***	***5.00 (<.001)***	***3.78 (<.001)***	***3.87 (<.001)***	**2.96 (<.01)**				
	WP		***5.86 (<.001)***	***6.80 (<.001)***	***5.21 (<.001)***	***4.84 (<.001)***	3.27 (<.01)	0.98 (0.33)	1.37 (0.17)	-0.08 (0.94)				
	P		***6.33 (<.001)***	***7.02 (<.001)***	***5.56 (<.001)***	***5.42 (<.001)***	***3.67 (<.001)***	1.57 (0.12)	1.93 (0.05)	0.51 (0.61)				
	S		***7.34 (<.001)***	***7.33 (<.001)***	***5.52 (<.001)***	***5.89 (<.001)***	***4.86 (<.001)***	**2.99 (<.01)**	**2.70 (<.01)**	0.43 (0.67)				
sloT5-large	I	***4.08 (<.001)***	***4.60 (<.001)***	***4.86 (<.001)***	***5.86 (<.001)***	***5.74 (<.001)***	***5.43 (<.001)***	***5.78 (<.001)***	***6.18 (<.001)***	***5.94 (<.001)***	***5.60 (<.001)***	***5.77 (<.001)***	***5.98 (<.001)***	***6.40 (<.001)***
	WP		***5.79 (<.001)***	***5.63 (<.001)***	***6.56 (<.001)***	***6.52 (<.001)***	***6.08 (<.001)***	***6.21 (<.001)***	***6.27 (<.001)***	***5.95 (<.001)***	***5.56 (<.001)***	***5.76 (<.001)***	***5.94 (<.001)***	***5.92 (<.001)***
	P		***5.78 (<.001)***	***5.92 (<.001)***	***6.76 (<.001)***	***6.69 (<.001)***	***6.26 (<.001)***	***6.33 (<.001)***	***6.33 (<.001)***	***6.01 (<.001)***	***5.70 (<.001)***	***5.94 (<.001)***	***6.14 (<.001)***	***6.20 (<.001)***
	S		***5.51 (<.001)***	***5.05 (<.001)***	***6.04 (<.001)***	***6.04 (<.001)***	***5.63 (<.001)***	***6.25 (<.001)***	***6.60 (<.001)***	***6.42 (<.001)***	***6.13 (<.001)***	***6.33 (<.001)***	***6.41 (<.001)***	***7.24 (<.001)***

Model	Layer / Input		13	14	15	16	17	18	19	20	21	22	23	24
sloT5-large (cont.)	I		***6.34 (<.001)***	***6.21 (<.001)***	***6.00 (<.001)***	***6.08 (<.001)***	***5.88 (<.001)***	***5.42 (<.001)***	***4.85 (<.001)***	***4.54 (<.001)***	***3.99 (<.001)***	***3.51 (<.001)***	3.20 (<.01)	3.27 (<.01)
	WP		***6.26 (<.001)***	***5.54 (<.001)***	***5.14 (<.001)***	***5.82 (<.001)***	***5.59 (<.001)***	***5.60 (<.001)***	***5.26 (<.001)***	***5.20 (<.001)***	***5.14 (<.001)***	***5.14 (<.001)***	***5.56 (<.001)***	2.71 (<.01)
	P		***6.50 (<.001)***	***5.67 (<.001)***	***5.30 (<.001)***	***5.95 (<.001)***	***5.62 (<.001)***	***5.63 (<.001)***	***5.32 (<.001)***	***5.27 (<.001)***	***5.14 (<.001)***	***5.26 (<.001)***	***5.66 (<.001)***	2.94 (<.01)
	S		***6.99 (<.001)***	***6.20 (<.001)***	***5.46 (<.001)***	***5.97 (<.001)***	***5.75 (<.001)***	***5.63 (<.001)***	***5.17 (<.001)***	***4.87 (<.001)***	***4.26 (<.001)***	***3.61 (<.001)***	***3.31 (<.001)***	1.43 (0.15)

Table 10: T-test results comparing cosine similarities of metaphoric and non-metaphoric noun-modifier constructions from KOMET 1.0. I = individual word inputs, WP = word pair input, P = complete phrase input, S = sentence input. Results with $p < .01$ in bold, $p < .001$ in bold italic. The highest difference in distributions per model is underlined

Model	Layer / Input	0	1	2	3	4	5	6	7	8	9	10	11	12
ELMo	I	***6.81 (<.001)***	***5.92 (<.001)***	***4.11 (<.001)***	***5.14 (<.001)***									
ELMo	WP		***5.48 (<.001)***	***4.91 (<.001)***	***5.53 (<.001)***									
ELMo	P		***6.20 (<.001)***	***5.43 (<.001)***	***6.09 (<.001)***									
ELMo	S		***5.55 (<.001)***	***4.26 (<.001)***	***5.03 (<.001)***									
SloBERTa	emb	***6.04 (<.001)***												
SloBERTa	I	***5.81 (<.001)***	***5.01 (<.001)***	***4.93 (<.001)***	1.91 (0.06)	1.28 (0.20)	0.63 (0.53)	0.38 (0.70)	0.59 (0.55)	0.40 (0.69)	0.25 (0.81)	0.18 (0.86)	-0.21 (0.83)	-0.54 (0.59)
SloBERTa	WP	***6.00 (<.001)***	***5.54 (<.001)***	***4.15 (<.001)***	1.48 (0.14)	0.34 (0.74)	-0.88 (0.38)	-1.44 (0.15)	-1.10 (0.27)	-1.30 (0.19)	-1.27 (0.20)	-1.19 (0.23)	-1.55 (0.12)	-2.17 (0.03)
SloBERTa	P	***5.88 (<.001)***	***5.28 (<.001)***	***4.78 (<.001)***	2.47 (0.01)	1.44 (0.15)	0.20 (0.85)	-0.20 (0.84)	0.04 (0.97)	-0.20 (0.84)	-0.14 (0.89)	-0.22 (0.83)	-0.40 (0.69)	-1.92 (0.05)
SloBERTa	S	***5.98 (<.001)***	***6.43 (<.001)***	***6.35 (<.001)***	***4.75 (<.001)***	***4.45 (<.001)***	***4.38 (<.001)***	***3.97 (<.001)***	***3.38 (<.001)***	***3.40 (<.001)***	**2.73 (<.01)**	**3.28 (<.01)**	**2.71 (<.01)**	-0.17 (0.87)
sloT5-small	I	***4.35 (<.001)***	***5.15 (<.001)***	***6.01 (<.001)***	***4.19 (<.001)***	***3.93 (<.001)***	***3.36 (<.001)***	2.34 (0.02)	2.49 (0.01)	***3.77 (<.001)***				
sloT5-small	WP		***5.92 (<.001)***	***6.47 (<.001)***	***3.83 (<.001)***	2.50 (0.01)	1.12 (0.26)	0.28 (0.78)	0.21 (0.84)	1.84 (0.07)				
sloT5-small	P		***6.16 (<.001)***	***6.79 (<.001)***	***4.99 (<.001)***	***4.17 (<.001)***	2.46 (0.01)	0.82 (0.41)	0.84 (0.40)	0.69 (0.49)				
sloT5-small	S		***7.32 (<.001)***	***6.76 (<.001)***	***5.85 (<.001)***	***6.31 (<.001)***	***5.33 (<.001)***	***4.16 (<.001)***	***4.06 (<.001)***	***3.60 (<.001)***				
sloT5-large	I	***6.38 (<.001)***	***5.87 (<.001)***	***5.28 (<.001)***	***6.18 (<.001)***	***6.01 (<.001)***	***5.53 (<.001)***	***5.84 (<.001)***	***5.76 (<.001)***	***5.51 (<.001)***	***5.34 (<.001)***	***5.41 (<.001)***	***5.45 (<.001)***	***5.90 (<.001)***
sloT5-large	WP		***7.03 (<.001)***	***5.62 (<.001)***	***6.44 (<.001)***	***6.41 (<.001)***	***5.89 (<.001)***	***5.94 (<.001)***	***5.41 (<.001)***	***4.99 (<.001)***	***5.04 (<.001)***	***5.06 (<.001)***	***5.34 (<.001)***	***5.11 (<.001)***
sloT5-large	P		***7.26 (<.001)***	***6.00 (<.001)***	***6.68 (<.001)***	***6.60 (<.001)***	***6.00 (<.001)***	***6.09 (<.001)***	***5.61 (<.001)***	***5.20 (<.001)***	***5.10 (<.001)***	***5.09 (<.001)***	***5.24 (<.001)***	***5.26 (<.001)***
sloT5-large	S		***6.92 (<.001)***	***5.28 (<.001)***	***6.28 (<.001)***	***6.28 (<.001)***	***5.64 (<.001)***	***6.19 (<.001)***	***6.20 (<.001)***	***5.87 (<.001)***	***5.65 (<.001)***	***5.67 (<.001)***	***5.90 (<.001)***	***6.41 (<.001)***
sloT5-large	(cont.)		13	14	15	16	17	18	19	20	21	22	23	24
sloT5-large	I		***5.64 (<.001)***	***5.34 (<.001)***	***5.31 (<.001)***	***5.21 (<.001)***	***5.08 (<.001)***	***5.00 (<.001)***	***4.78 (<.001)***	***4.57 (<.001)***	***4.42 (<.001)***	***3.82 (<.001)***	***3.48 (<.001)***	***3.64 (<.001)***
sloT5-large	WP		***4.94 (<.001)***	***4.94 (<.001)***	***4.96 (<.001)***	***5.10 (<.001)***	***5.14 (<.001)***	***5.05 (<.001)***	***5.10 (<.001)***	***4.87 (<.001)***	***4.71 (<.001)***	***3.94 (<.001)***	***3.84 (<.001)***	***5.90 (<.001)***
sloT5-large	P		***5.21 (<.001)***	***4.74 (<.001)***	***4.70 (<.001)***	***4.93 (<.001)***	***4.77 (<.001)***	***4.53 (<.001)***	***4.69 (<.001)***	***4.55 (<.001)***	***4.43 (<.001)***	***3.64 (<.001)***	***3.83 (<.001)***	***4.41 (<.001)***
sloT5-large	S		***6.32 (<.001)***	***6.80 (<.001)***	***6.80 (<.001)***	***7.04 (<.001)***	***7.08 (<.001)***	***7.05 (<.001)***	***6.85 (<.001)***	***6.73 (<.001)***	***6.97 (<.001)***	***6.26 (<.001)***	***5.94 (<.001)***	***5.02 (<.001)***

type and layer are statistically different. The only exception to this is the final layer (24), S input, for amod constructions. The largest difference in distribution for amod constructions is observed on layer 12, S input ($t=7.24$, $p<.001$), and for nmod constructions on layer 1, P input ($t=7.26$, $p<.001$). This is the only case where the highest t-score is achieved from contextual embeddings generated from a phrase input to the model.

In Figure 1 and Figure 2, we visualize the effect sizes in terms of Cohen's *d* for each pair of distributions per model and layer. In general, we can observe that the effect of cosine similarity difference in amod constructions is lower than that in nmod constructions. The largest overall effects can be observed in embeddings obtained from fT_EMBEDDIA, those obtained from the ELMo input (0th) layer, and those obtained from the sentence input to sloT5-small and sloT5-large, all achieving around $d=0.4$, which can be interpreted as small to medium. The effect sizes for cosine similarities obtained from SloBERTa are on the smaller side, especially in the case of adjective modifier constructions.

Among the static models, w2v_ELMo embeddings capture almost no difference in the two distributions, with a $d=0.11$ and $d=0.13$.

	d
w2v_ELMo	0.13
fT_EMBEDDIA	0.39
fT_CLARIN	0.36

ELMo

Input \ Layer	0	1	2	3
I	0.38	0.31	0.27	0.33
WP	0.38	0.31	0.23	0.29
P	0.38	0.30	0.23	0.28
S	0.38	0.21	0.29	0.30

sloBERTa

Input \ Layer	0	1	2	3	4	5	6	7	8	9	10	11	12
emb	0.20												
I	0.26	0.18	0.13	0.04	0.08	0.11	0.09	0.10	0.07	0.09	0.11	0.13	0.12
WP	0.25	0.26	0.22	0.03	0.08	0.09	0.09	0.06	0.04	0.03	0.17	0.22	0.18
P	0.25	0.25	0.24	0.01	0.05	0.04	0.03	0.02	0.01	0.03	0.08	0.14	0.12
S	0.24	0.25	0.27	0.10	0.05	0.07	0.06	0.06	0.05	0,00	0.02	0.07	0.05

sloT5-small

Input \ Layer	0	1	2	3	4	5	6	7	8
I	0.23	0.28	0.34	0.27	0.27	0.26	0.20	0.20	0.15
WP	0.23	0.30	0.35	0.27	0.25	0.17	0.05	0.07	0.00
P	0.23	0.33	0.37	0.29	0.28	0.19	0.08	0.10	0.03
S	0.23	0.38	0.38	0.29	0.31	0.25	0.16	0.14	0.02

sloT5-large

Input \ Layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
I	0.21	0.24	0.25	0.31	0.30	0.28	0.3	0.32	0.31	0.29	0.30	0.31	0.33	0.33	0.32	0.31	0.32	0.31	0.28	0.25	0.24	0.21	0.18	0.17	0.17
WP	0.21	0.30	0.29	0.34	0.34	0.32	0.32	0.33	0.31	0.29	0.30	0.31	0.31	0.33	0.29	0.27	0.30	0.29	0.29	0.27	0.27	0.27	0.27	0.29	0.14
P	0.21	0.30	0.31	0.35	0.35	0.33	0.33	0.33	0.31	0.30	0.31	0.32	0.32	0.34	0.30	0.28	0.31	0.29	0.29	0.28	0.27	0.27	0.27	0.29	0.15
S	0.21	0.29	0.26	0.31	0.31	0.29	0.33	0.34	0.33	0.32	0.33	0.33	0.38	0.36	0.32	0.28	0.31	0.30	0.29	0.27	0.25	0.22	0.19	0.17	0.07

Figure 1: Effect sizes for amod constructions in terms of Cohen's *d*. emb = embedding layer, I = individual word inputs, WP = word pair input, P = complete phrase input, S = sentence input. Dark purple = larger effect, light yellow = smaller effect

	d
w2v_ELMo	0.11
fT_EMBEDDIA	0.43
fT_CLARIN	0.34

ELMo

Input \ Layer	0	1	2	3
I	0.40	0.34	0.24	0.30
WP	0.40	0.32	0.29	0.32
P	0.40	0.36	0.32	0.35
S	0.40	0.32	0.25	0.29

sloBERTa

Input \ Layer	0	1	2	3	4	5	6	7	8	9	10	11	12
emb	0.35												
I	0.34	0.29	0.29	0.11	0.07	0.04	0.02	0.03	0.02	0.01	0.01	0.01	0.03
WP	0.35	0.32	0.24	0.09	0.02	0.05	0.08	0.06	0.08	0.07	0.07	0.09	0.13
P	0.34	0.31	0.28	0.14	0.08	0.01	0.01	0,00	0.01	0.01	0.01	0.02	0.11
S	0.35	0.37	0.37	0.28	0.26	0.25	0.23	0.20	0.20	0.16	0.19	0.16	0.01

sloT5-small

Input \ Layer	0	1	2	3	4	5	6	7	8
I	0.25	0.30	0.35	0.24	0.23	0.20	0.14	0.14	0.22
WP	0.25	0.34	0.38	0.22	0.15	0.06	0.02	0.01	0.11
P	0.25	0.36	0.40	0.29	0.24	0.14	0.05	0.05	0.04
S	0.25	0.43	0.39	0.34	0.37	0.31	0.24	0.24	0.21

sloT5-large

Input \ Layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
I	0.37	0.34	0.31	0.36	0.35	0.32	0.34	0.34	0.32	0.31	0.32	0.32	0.34	0.33	0.31	0.31	0.30	0.30	0.29	0.28	0.27	0.26	0.22	0.20	0.21
WP	0.37	0.41	0.33	0.37	0.37	0.34	0.35	0.32	0.29	0.29	0.29	0.31	0.30	0.29	0.29	0.29	0.30	0.30	0.29	0.3	0.28	0.27	0.23	0.22	0.34
P	0.37	0.42	0.35	0.39	0.38	0.35	0.35	0.33	0.30	0.30	0.30	0.30	0.31	0.30	0.28	0.27	0.29	0.28	0.26	0.27	0.26	0.26	0.21	0.22	0.26
S	0.37	0.40	0.31	0.37	0.37	0.33	0.36	0.36	0.34	0.33	0.33	0.34	0.37	0.37	0.40	0.40	0.41	0.41	0.41	0.40	0.39	0.41	0.36	0.35	0.29

Figure 2: Effect sizes for nmod constructions in terms of Cohen's *d*. emb = embedding layer, I = individual word inputs, WP = word pair input, P = complete phrase input, S = sentence input. Dark purple = larger effect, light yellow = smaller effect

Both fT_EMBEDDIA and fT_CLARIN show a small to medium effect, with fT_EMBEDDIA as the best static model for both types of constructions. Comparing the contextual models, we can see a trend for ELMo and SloBERTa, where embeddings from the input and the lower hidden layers consistently better disambiguate metaphoric from non-metaphoric constructions than the subsequent upper layers. In sloT5, a similar observation can be made for the lower layers (not including the input layer). The largest effects for amod and nmod constructions can be observed in embeddings from layer 1 and layer 2. In the larger sloT5 variant, the effect size trend differs by construction. Whereas the late-lower and middle layers (3–13) are better for amod constructions, nmod constructions are better disambiguated in lower and upper layers (0–4; 12–21). SloT5-large also has the overall most stable effect of embedding dissimilarity throughout the layers.

While the largest effect is still observed in one of the static fastText embeddings for both amod constructions (d = .39) and nmod constructions (d = .43), contextual embedding methods are only marginally behind. The effect can be interpreted as small to medium, as there is still a large overlap between the cosine similarities of metaphoric and non-metaphoric pairs. However, on average, non-metaphoric word pairs retain a higher cosine similarity in an overwhelming majority of the experimental settings.

The overall larger effect sizes for nmod constructions could be due to the larger samples, or inherent semantic and syntactic differences between nouns and adjectives. Nouns typically exhibit greater syntactic flexibility, functioning in various syntactic roles. In contrast, adjectives primarily serve as noun modifiers, with which they classify the noun into a domain, attribute characteristics or express ownership (Vidovič-Muha 1978). In cognitive grammar (Langacker 1990), nouns correspond to things and are considered conceptually autonomous, while adjectives express relations and are considered conceptually dependent. Adjectives are thus often also semantically underspecified and rely on accompanying nouns for complete interpretation (Paradis 2000). Both these "dependencies", syntactic and semantic, may also translate into the vector space representations, potentially making adjectives closer to nouns in general.

5 ANALYSIS OF TOP AND BOTTOM EXAMPLES

In this section, we analyze metaphoric and literal examples from both tails of the distributions, namely phrases with the highest and the lowest cosine similarity. The first part of the analysis concerns the relation between collocation strength and cosine similarity. In the second part of the analysis, we manually study the examples to gain insights into why some pairs appear at the left and others at the right tail of the cosine similarity distribution. We analyze examples from all static embeddings, whereas for each contextual model, we limit the analysis to the layer with the largest effect size in terms of Cohen's *d* from the previous section. We sample ten constructions from the top, that is, those with the highest cosine similarities, and ten from the bottom, that is, those with the lowest cosine similarities, separately for nmod and amod constructions.

Quantitative analysis and collocability 5.1

Out of the 240 sampled top and bottom constructions from static models, 185 are unique in total. Seven of those are shared across the three models, 38 appear in two, and 140 are unique to just one model. Altogether 83 of them are unique to static models, meaning they do not appear in the tails of the contextual models.

Zooming in on only the top and bottom ten metaphoric noun-modifier constructions (Table 11) from fT_EMBEDDIA, the static model with the largest effect size, we can already observe a pattern. Constructions which contain words most similar in terms of cosine similarity are indeed frequent expressions in language, such

	Slovene	English
most similar	**POGLED** V PRIHODNOST	[a] **LOOK** INTO [the] FUTURE
	razjasnitve vseh okoliščin	**clarification** [of] all circumstances
	izrazna **večplastnost***	expressive **multilayeredness***
	žrtve nasilja	**victims** [of] violence
	dosego cilja	**achieving** [a/the] goal
	nebo **vpijoča***	sky **screaming** [=*egregious, obvious*]*
	premagovanju ovir	**conquering** [of] obstacles
	vsem **vpletenim***	all **woven into** [=*all involved*]*
	magnetizem privlačnosti	**magnetism** [of] attraction
	krogu kvalifikacij	**circle** [of] qualifications [=*qualification round*]
least similar	**lov** za morebitnimi poslušalci	**hunt** for potential listeners
	njim **povezana***	him **connected** [=*related to him/it*]
	gori podatkov	**mountain** [of] data
	na primer žalujemo*	**on**[=*for*] example [we] grieve*
	okviru akrobatskega smučanja	**frame** [of] freestyle skiing
	srce vsakega muzeja	**heart** [of] every museum
	delničar cele **vrste**	shareholder [of a] whole **row** [=*range*]
	nos za prihodnje potrebe	**nose** for future needs
	jezik moči	**language** [of] power
	poslopja moči	**edifices** [of] power

Table 11: Ten most similar and least similar word pairs in nmod constructions in terms of cosine similarity in static fT_EMBEDDIA embeddings. The metaphoric word is in bold, an asterisk (*) indicates erroneous syntactic dependency annotation

as *dosego cilja* 'achieving [a/the] goal', *žrtve nasilja* 'victims of violence', *krog kvalifikacij* 'qualification round'. In the set of constructions with least similar constituents in terms of cosine similarity, we observe less frequent and more expressive phrases, such as *lov za morebitnimi poslušalci* 'hunt for potential listeners', *gori podatkov* 'mountain of data', *nos za prihodnje potrebe* 'nose for future needs'. This indicates that the cosine similarity of the constituents of metaphoric constructions correlates with metaphor conventionality/novelty.

Contextual embedding models differ somewhat in which constructions contain the most and least similar words in terms of cosine similarity but they share quite a few of the examples. Altogether, there are 256 unique examples of all the sampled 320. Three of those are shared across the four models; eight examples are shared between all but one model, 39 appear in two models. A total of 206 examples uniquely appear in the top/bottom ten of one of the models. This suggests that, although the general mechanisms are similar, not only is word meaning encoded differently or in different locations depending on the model, embeddings contain very different information.

Figure 3 demonstrates the collocability of constructions found in the list of collocations (those not appearing among the collocations are excluded). The collocation strength is measured with logDice score (Rychlý 2008), which takes into account the frequency of a word pair relative to the frequencies of the individual words. We can observe the top ten constructions mostly rank at the higher end of collocation strength, while the bottom ten rank at the lower end. This is evidently the case for static embeddings and SloBERTa, where the top ten examples of each of the categories are on average much more collocationally bound than the bottom ten examples. In ELMo, this is true for non-metaphoric amod constructions and both metaphoric and non-metaphoric nmod constructions. However, no clear difference can be discerned for metaphoric amod constructions (the first pair of bars is very close together, with a similar number of collocations sampled). In the sloT5 variants, the relationship between cosine similarity and collocation strength is even less straightforward. In the case of non-metaphoric amod constructions in sloT5-small (second pair of bars), a reverse trend is seen, meaning examples from the bottom of the distribution have, on average, greater collocation strength. In the case

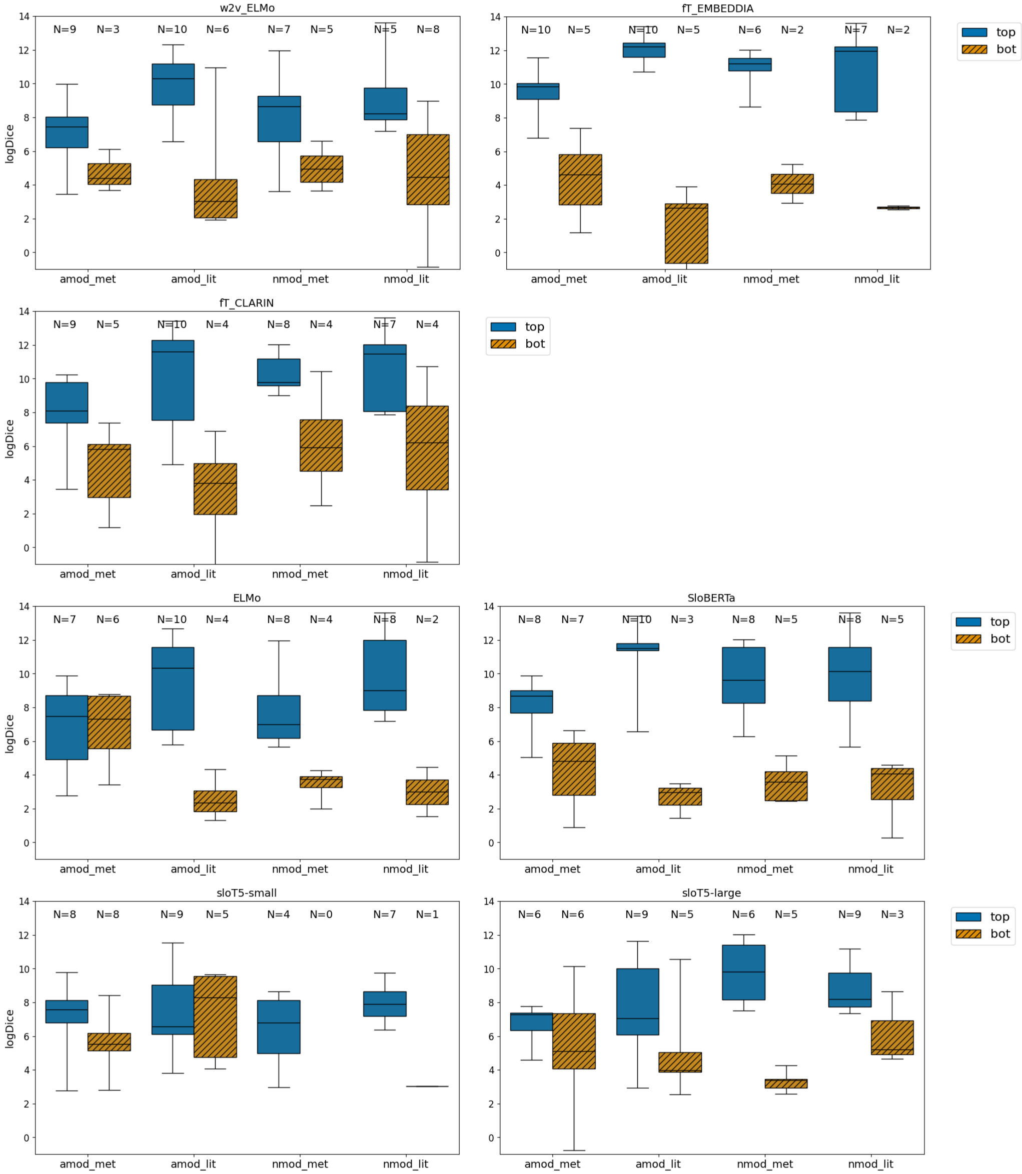

Figure 3: Collocation strengths in terms of the logDice coefficient of the ten most similar (blue) and ten least similar (orange) word pairs per construction type

of sloT5-large, metaphoric amod constructions (the first pair of bars) also range in the same area of collocability.

To further explore the relation between collocation strength and cosine similarity, we generate scatter plots featuring these two variables in Figures 4 and 5. Note that the plots only include the sampled topmost and bottommost constructions, not the entire dataset. Still, we can observe a tendency by which weaker collocation strength is correlated with lower cosine similarity (lower left quadrant), and greater collocation strength with higher cosine similarity (upper right quadrant), most evidently so for fT_EMBEDDIA static embeddings. In sloT5-small, however, we observe no such trend.

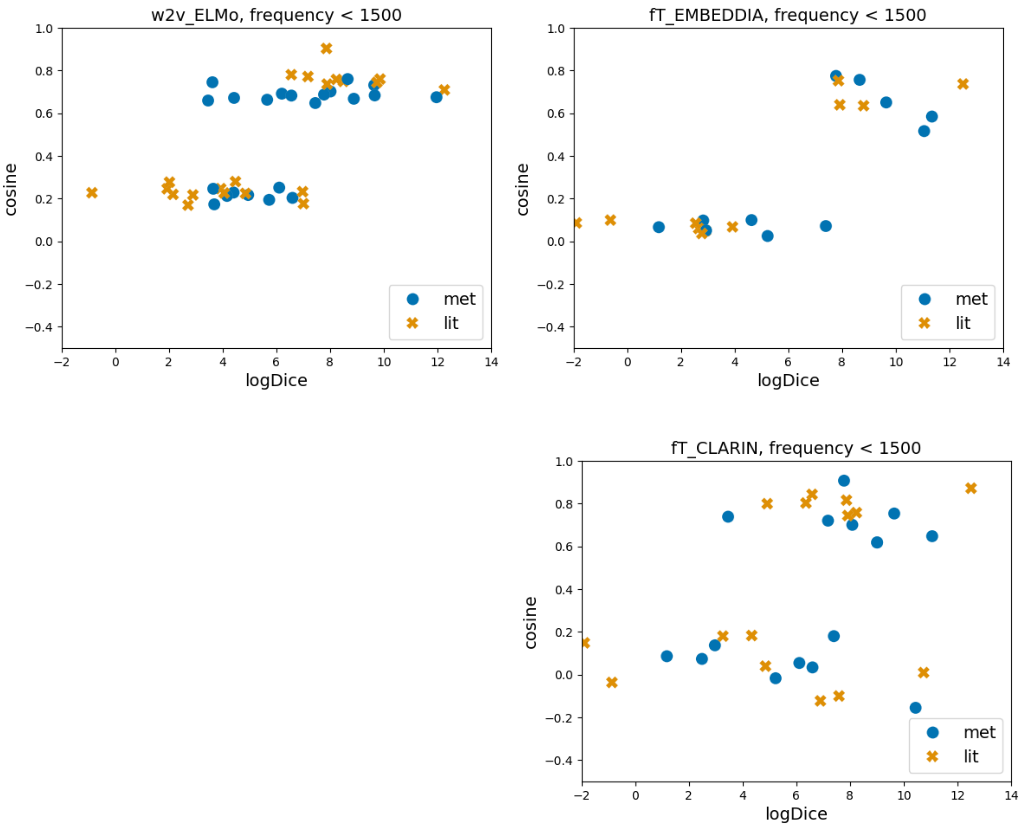

Figure 4: Collocation strengths in terms of the logDice coefficient vs. cosine similarity of metaphoric and non-metaphoric constructions in static embeddings

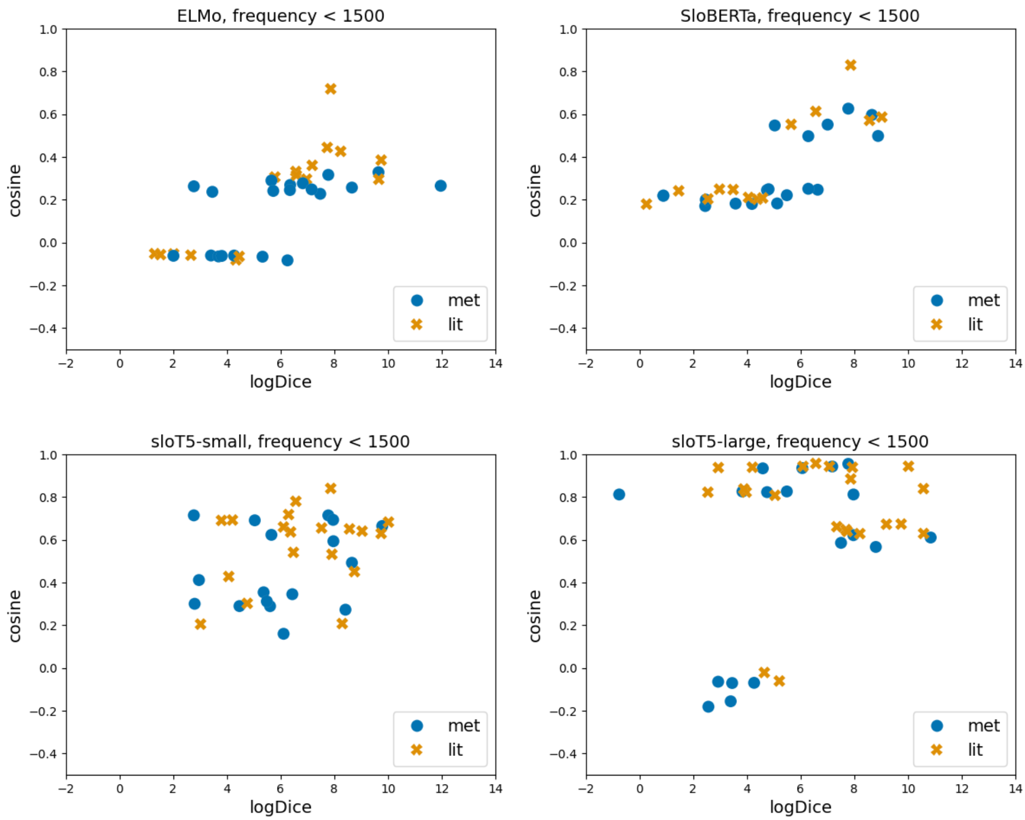

Figure 5: Collocation strengths in terms of the logDice coefficient vs. cosine similarity of metaphoric and non-metaphoric constructions in contextual embeddings

5.2 *Manual analysis*

In this section, we manually analyze both metaphoric and non-metaphoric constructions with the highest and lowest cosine similarities between constituents, for both types of constructions. We deduplicate the examples collected from all the models, resulting in altogether 304 unique constructions: 84 of those are bottommost metaphoric, 69 topmost metaphoric, 81 bottommost literal, and 70 topmost literal constructions. In the process of manual analysis, we annotate them with the following ten non-exclusive observation categories:

1. `collocations`: the constituents appear in the list of collocations,

2. `possible literal`: the construction containing a metaphor could be read literally, i.e., it is not metaphoric on the level of the relation between the two words without additional context,
3. `possible metaphor`: a word in the literal construction could be considered metaphoric,
4. `annotation error`: the construction in question was not properly linguistically parsed or annotated,
5. `uppercased word`: the construction contains an uppercased word,
6. `foreign word`: one or both of the constituents are foreign words,
7. `long-range dependency`: the constituents are many words apart, i.e., with two or more tokens in between,
8. `capitals:` all the words are in capital letters,
9. `similar domains`: the words come from similar semantic domains,
10. both `metaphoric/idiom`: both words may be considered metaphoric or form an idiomatic expression.

The results of the manual analysis are demonstrated in Table 12. We can observe all of the sets have a large number of examples appearing in the list of collocations, most evidently in the case of metaphoric (58/69 or 84.1%) and non-metaphoric (70/83 or 84.3%) constructions with most similar constituents. In the set of constructions with least similar constituents, we observe quite a lot of examples that were extracted based on erroneous linguistic annotations (21 metaphors and 20 non-metaphoric constructions). An example is *okviru, in številne* 'framework, and numerous', where the linguistic annotation marks *številne* as an adjective modifier of *okviru*, however, the constituents actually belong to two separate clauses, and *številne* modifies a different noun in the sentence. We also see many cases where one of the words is uppercased. Another possible cause for low-cosine-similarity examples are long-range dependencies, i.e., cases where the two words supposedly participating in a syntactic relation are separated with many words in between. An instance of a low-similarity literal example is *ena izmed največjih pomorskih katastrof* 'one of the largest maritime catastrophes', although we also identify 'katastrof' as a possible metaphor. In fact, among the ten literal

Table 12: Combined analysis of metaphoric and non-metaphoric constructions with most (top) and least (bottom) similar constituents. Metaphoric words in bold. Words identified as potentially metaphoric in the analysis underlined

Observation	**bot_met**	top_met	bot_lit	top_lit	Example
collocations	51	58	47	70	*nogometne* ***poti*** 'football **path**'
possible literal	23	20	–	–	*širokih pripovednih* ***lokih*** 'wide narrative **arcs**'
possible metaphor	–	–	30	4	*zvesta filmu* '**loyal** [to the] film'
annotation error	21	7	20	2	***okviru****, in številne* '**framework**, and numerous'
uppercased word	11	2	26	6	*SP* ***divizije*** 'world championship **division**'
long–range dependency	14	4	12	2	*ena izmed največjih pomorskih* ***katastrof*** 'one of the largest maritime **catastrophes**'
similar domains	–	1	–	–	***toča*** *storžev* '**hail** of acorns'
both words metaphoric/idiom	1	2	–		***toča*** *storžev* '**hail** of acorns'
metaphor annotation error	–	1	–	–	*letih delovne* ***dobe*** 'years of working **age**'
capitals	–	1	–		***POGLED*** *V PRIHODNOST* '[a] **LOOK** INTO [the] FUTURE'
foreign word(s)	1	–	3	1	*hat trick*
Total analysed	111	69	108	83	

constructions with the lowest cosine similarity similar constituents, we find as many as 30 examples where at least one of the words could be regarded as metaphoric, for example, *zvesta* in *zvesta filmu* 'loyal [to the] film'. With regards to the metaphoric constructions with the lowest cosine similarity between constituents, we observe 15 cases where the construction could be read literally. An example of a possible literal read is *širokih pripovednih lokih* 'wide narrative arcs', where only based on the two main constituents (wide, arc), the phrase could be interpreted literally. However, when considering the inner part of the construction, with *pripovednih* 'narrative' as the adjective modifier of *lokih* 'arcs', the latter is clearly metaphorically used. The cosine similarity of the constituents of this inner amod relation is also much lower (0.139 compared to 0.269 in ELMo).

Although these results are not the most insightful for metaphor identification, they nevertheless provide insights into other corollaries of a low cosine similarity metric: different word capitalization, foreign words, collocability, and different syntactic relations. As the analysis shows, an extremely low cosine similarity metric could indicate wrong data annotations of both linguistic structures and metaphoricity. Namely, by looking at the supposedly literal examples with the lowest similarity between constituents, we found quite a few metaphors. Some of those (e.g., ***zvesta*** *filmu* '**loyal** to [the] film', *pravni* ***aparat*** 'legal **apparatus**', ***tenkočutno*** *razstavo* 'sensitive/perceptive exhibition'), as more creative and clear metaphors, are even more interesting to analyze than some of the high-similarity metaphoric cases which usually correspond to established, conventionalized metaphors (e.g., ***pekoča*** *bolečina* '**stinging** pain', ***dosego*** *cilja* '**achieving** a goal', *ljubezensko* ***romanco*** 'love **romance**'). On the other hand, cosine similarity also seems to encode the differences in word capitalization. We observe many mismatches in terms of lowercased/uppercased words at the bottom of both metaphoric and non-metaphoric distributions, while an all-capitalized example ***POGLED*** *V PRIHODNOST* 'a **look** into the future' appears in the set of constructions with most similar relation constituents.

In line with the results of the manual annotation which revealed potential annotation errors, an expert linguist re-annotated a random sample following the general principles of the metaphor identification procedure MIPVU. The re-annotation was conducted on four randomly selected texts from KOMET or 4,416 tokens. The inter-annotator agreement calculated with Cohen's (1960) kappa reached κ =0.62, which may be considered substantial (Landis and Koch 1977). However, this rate is much lower than the agreement rate for other metaphor annotation campaigns which were based on formalized procedures and involved several rounds of discussion, deliberation, and (re-)annotation (see Steen 2010 and Nacey *et al.* 2019). To ensure better quality of the data and provide stronger support for our results, the metaphor corpus should be re-annotated by first adopting a fully formalized procedure, involving several expert annotators, and performing many rounds of discussion and annotation.

CONCLUSION 6

In this article, we studied word embeddings from static and contextual models in their ability to represent semantic information. The study extends previous work on the intrinsic analysis of language models by zooming in on the representation of basic meaning, and, consequently, its use for metaphor identification in Slovene. Based on the hypothesis that words in relation-level metaphors originate from different domains, we investigated this semantic incongruity between words with the help of cosine similarity. Our results mostly confirm this hypothesis, namely non-metaphoric word pairs are on average placed closer together in terms of cosine similarity than metaphoric combinations. However, not all language models were able to capture this difference, and the effect of metaphoricity on cosine similarity also greatly differs by model and layer. While all monolingual models differentiate between metaphoric and non-metaphoric constructions, no statistically significant difference could be derived from the cosine similarities obtained from multilingual representations. This finding is in line with previous studies of internal model representations (Vulić *et al.* 2020) showing monolingual models perform better on downstream lexico-semantic tasks. We believe one of the reasons for the inferior results in our study may be simply due to a different tokenization strategy in multilingual models (Rust *et al.* 2021). A multilingual tokenizer results in a much more fragmentary space of token embeddings to cover multiple languages and thus fails to capture relevant semantics in these small fragments. Furthermore, comparing static and contextual embeddings of monolingual models, our results indicate contextual embeddings are comparable to but not above static embeddings.

The results also mostly confirm previous findings, namely word embeddings become increasingly more contextualized in the upper layers of the model, and type-level representations are contained in the lower layers. An exception to that was the largest model in our study (sloT5-large), where the middle layers seemed to include the most relevant information to disambiguate metaphoric from non-metaphoric constructions. The results also show different patterns for noun-modifier and adjective-modifier constructions, implying a non-negligible effect of syntax. In terms of the different input contexts to contextual models, we find that the sentence context performs the best, although results from other input types are comparable.

In the last part of our analysis, we observe an effect of collocation strength on cosine similarity. This finding has two key implications. On the one hand, this suggests the latter could be exploited to disambiguate novel and conventional metaphors. This is in line with Li *et al.* (2023a), who note that word embeddings do not necessarily reflect a single "basic" meaning, but rather what they describe as "aggregated meaning" – a blend of all word senses, weighted by their frequency in actual language use. In the case of highly conventional metaphors, this aggregated meaning may, in fact, predominantly reflect the metaphorical sense. On the other hand, co-occurrence could influence cosine similarity more than the basic semantic meaning or semantic similarity of two words. Beyond collocation, we identify additional factors that significantly affect cosine similarity, including word capitalization and long-range syntactic dependencies. Our findings are consistent with previous research (e.g., Zhou *et al.* 2022) that questions the reliability of cosine similarity as it is heavily influenced by the frequency of the tokens in the training data.

In conclusion, our findings suggest both static and contextual embeddings incorporate semantic information about the basic meaning of words, relevant for relation-level metaphor detection via cosine similarity. However, the effectiveness and location of that information were highly dependent both on the type of construction and model involved. Moreover, analyses also reveal other factors than just the semantics of words to have a potential impact on cosine similarity, which we plan to explore and remediate in future work. While cosine similarity may not reliably indicate the presence of metaphor, our results suggest it can still provide insight into the degree of conventionality of a particular metaphor.

7 LIMITATIONS AND FUTURE WORK

Our study was focused on adjective-noun and noun-noun phrases, which is why we cannot draw conclusions about the usefulness of this approach for the representation or identification of metaphors or basic meanings in other types of constructions. Secondly, the corpus of metaphors used as a dataset in this work has not yet been sufficiently validated, as it was only annotated by one person. Although a sample of data was re-annotated, the inter-annotator agreement rate did not reach that of high-quality datasets, underscoring the need for re-annotation and a formalized annotation procedure, adapted to the Slovene linguistic and extralinguistic context. Our experiments also rely on automatic linguistic annotations of syntactic dependencies and parts of speech from an older linguistic processing pipeline. Although we uncovered annotation errors through manual analysis, in future work, linguistic structures should be re-annotated with the state-of-the-art processing pipeline. Our experiments explored the cosine similarity metric as a measure of semantic similarity and incongruity. Future work could explore other similarity/distance metrics and perhaps reach other findings. More importantly, several other methods could be used to probe for semantic or metaphoricity information contained in word embeddings but which are unfortunately out of scope. Due to length limitations, we do not yet directly conduct the identification of metaphors and thus do not compare to previous approaches. Lastly, our study is limited to Slovene and does not imply the same observed patterns to be true for embedding models for other languages.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support from the Slovene Research Agency (ARRS) under the research core funding P6-0215. The authors also gratefully acknowledge the HPC RIVR consortium (`www.hpc-rivr.si`) and EuroHPC JU (`eurohpc-ju.europa.eu`) for funding this research by providing computing resources of the Vega HPC System at the Institute of Information Science (`www.izum.si`).

REFERENCES

Kat R. Agres, Stephen McGregor, Karolina Rataj, Matthew Purver, and Geraint A. Wiggins (2016), Modeling metaphor perception with distributional semantics vector space models, in *Proceedings of the ESSLLI Workshop on Computational Creativity, Concept Invention, and General Intelligence (C3GI)*, pp. 1–14.

Špela Antloga (2020a), Korpus metafor KOMET 1.0, in Darja Fišer and Tomaž Erjavec, editors, *Proceedings of the Conference on Language Technologies and Digital Humanities*, pp. 167–170, http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Student_Antloga_Korpus-metafor-KOMET-1.0.pdf.

Špela Antloga (2020b), Metaphor corpus KOMET 1.0, http://hdl.handle.net/11356/1293, Slovenian language resource repository CLARIN.SI.

Špela Antloga and Gregor Donaj (2022), Corpus of metaphorical expressions in spoken Slovene language G-KOMET 1.0, https://www.clarin.si/repository/xmlui/handle/11356/1490.

Mateusz Babieno, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki (2022), MIss RoBERTa WiLDe: Metaphor Identification Using Masked Language Model with Wiktionary Lexical Definitions, *Applied Sciences*, 12(4), doi:10.3390/app12042081, https://www.mdpi.com/2076-3417/12/4/2081.

Andreas Blank (1999), Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change, in Andreas Blank and Peter Koch, editors, *Historical Semantics and Cognition*, pp. 61–90, De Gruyter Mouton, doi:10.1515/9783110804195.61.

Amber Boeynaems, Christian Burgers, Elly A. Konijn, and Gerard J. Steen (2017), The effects of metaphorical framing on political persuasion: A systematic literature review, *Metaphor and Symbol*, 32(2):118–134, doi:10.1080/10926488.2017.1297623.

Mojca Brglez (2023), Dispersing the clouds of doubt: can cosine similarity of word embeddings help identify relation-level metaphors in Slovene?, in Jakub Piskorski, Michał Marcińczuk, Preslav Nakov, Maciej Ogrodniczuk, Senja Pollak, Pavel Přibáň, Piotr Rybak, Josef Steinberger, and Roman Yangarber, editors, *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pp. 61–69, Association for Computational Linguistics, Dubrovnik, Croatia, doi:10.18653/v1/2023.bsnlp-1.8, https://aclanthology.org/2023.bsnlp-1.8.

Mojca Brglez, Senja Pollak, and Špela Vintar (2021), Simple discovery of COVID IS WAR metaphors using word embeddings, in Dunja Mladenić and

Marko GROBELNIK, editors, *Odkrivanje znanja in podatkovna skladišča - SiKDD: 4 October 2021, Ljubljana, Slovenia*, pp. 37–40, Institut “Jožef Stefan”.

Mojca BRGLEZ, Omnia ZAYED, and Paul BUITELAAR (2025), TCMeta: a multilingual dataset of COVID tweets for relation-level metaphor analysis, *Language Resources and Evaluation*, 59:437–475, doi:10.1007/s10579-024-09725-z.

Laura BURDICK, Jonathan K. KUMMERFELD, and Rada MIHALCEA (2022), Using paraphrases to study properties of contextual embeddings, in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4558–4568, Association for Computational Linguistics, Seattle, WA, USA, doi:10.18653/v1/2022.naacl-main.338, `https://aclanthology.org/2022.naacl-main.338`.

Christian BURGERS, Elly KONIJN, and Gerard STEEN (2016), Figurative framing: Shaping public discourse through metaphor, hyperbole, and irony, *Communication Theory*, 26:410–430, doi:10.1111/comt.12096.

Lynne CAMERON (2003), *Metaphor in educational discourse*, Advances in Applied Linguistics, Bloomsbury Publishing, ISBN 9781441175649.

Jonathan CHARTERIS-BLACK (2004), *Corpus approaches to critical metaphor analysis*, Palgrave Macmillan UK.

Minjin CHOI, Sunkyung LEE, Eunseong CHOI, Heesoo PARK, Junhyuk LEE, Dongwon LEE, and Jongwuk LEE (2021), MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1763–1773, Association for Computational Linguistics, Online, doi:10.18653/v1/2021.naacl-main.141, `https://aclanthology.org/2021.naacl-main.141`.

Jacob COHEN (1960), A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20(1):37–46, doi:10.1177/001316446002000104.

Jacob COHEN (2013), *Statistical power analysis for the behavioral sciences*, Taylor & Francis, ISBN 9781134742707.

Alexis CONNEAU, Kartikay KHANDELWAL, Naman GOYAL, Vishrav CHAUDHARY, Guillaume WENZEK, Francisco GUZMÁN, Edouard GRAVE, Myle OTT, Luke ZETTLEMOYER, and Veselin STOYANOV (2020), Unsupervised cross-lingual representation learning at scale, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.acl-main.747, `https://aclanthology.org/2020.acl-main.747`.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch (2019), Unsupervised compositionality prediction of nominal compounds, *Computational Linguistics*, 45(1):1–57, doi:10.1162/coli_a_00341, https://aclanthology.org/J19-1001.

Alan Cruse (2000), *Meaning in language: An introduction to semantics and pragmatics*, Oxford University Press, ISBN 9780198700104.

Alan Cruse (2006), *A glossary of semantics and pragmatics*, Edinburgh University Press, ISBN 9780748626892, doi:10.1515/9780748626892.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, in *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. 4171–4186, Association for Computational Linguistics, Minneapolis, MN, USA, doi:10.18653/v1/N19-1423, https://aclanthology.org/N19-1423.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych (2018), Weeding out conventionalized metaphors: A corpus of novel metaphor annotations, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1424, Association for Computational Linguistics, Brussels, Belgium, doi:10.18653/v1/D18-1171, https://aclanthology.org/D18-1171.

Mohamad Elzohbi and Richard Zhao (2024), ContrastWSD: Enhancing metaphor detection with word sense disambiguation following the metaphor identification procedure, in Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 3907–3915, ELRA and ICCL, Torino, Italy, https://aclanthology.org/2024.lrec-main.346.

Kawin Ethayarajh (2019), How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Association for Computational Linguistics, Hong Kong, China, doi:10.18653/v1/D19-1006, https://aclanthology.org/D19-1006.

Polona Gantar, Lut Colman, Carla Parra Escartín, and Héctor Martínez Alonso (2018), Multiword expressions: Between lexicography and NLP, *International Journal of Lexicography*, 32(2):138–162, doi:10.1093/ijl/ecy012.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio (2021), Probing for idiomaticity in vector space models, in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3551–3564, Association for Computational Linguistics, Online, doi:10.18653/v1/2021.eacl-main.310, `https://aclanthology.org/2021.eacl-main.310`.

Andrew GOATLY (2011), Metaphor as resource for the conceptualisation and expression of emotion, in *Affective computing and sentiment analysis: Emotion, metaphor and terminology*, pp. 13–25, Springer.

John HEWITT and Christopher D. MANNING (2019), A structural probe for finding syntax in word representations, in *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. 4129–4138, Association for Computational Linguistics, Minneapolis, MN, USA, doi:10.18653/v1/N19-1419, `https://aclanthology.org/N19-1419`.

Walter KINTSCH (2000), Metaphor comprehension: A computational theory, *Psychonomic Bulletin & Review*, 7:257–266, doi:10.3758/BF03212981.

Matej KLEMEN and Marko ROBNIK-ŠIKONJA (2023), Neural metaphor detection for Slovene, in *Selected papers from the CLARIN Annual Conference 2022, Linköping Electronic Conference Proceedings 198*, pp. 77–89, doi:10.3384/ecp198008.

Zoltán KÖVECSES (2020), *Extended conceptual metaphor theory*, Cambridge University Press, doi:10.1017/9781108859127.

Simon KREK, Tomaž ERJAVEC, Andraž REPAR, Jaka ČIBEJ, Špela ARHAR HOLDT, Polona GANTAR, Iztok KOSEM, Marko ROBNIK-ŠIKONJA, Nikola LJUBEŠIĆ, Kaja DOBROVOLJC, Cyprian LASKOWSKI, Miha GRČAR, Peter HOLOZAN, Simon ŠUSTER, Vojko GORJANC, Marko STABEJ, and Nataša LOGAR (2019), Corpus of written standard Slovene Gigafida 2.0, `http://hdl.handle.net/11356/1320`, Slovenian language resource repository CLARIN.SI.

Simon KREK, Polona GANTAR, Iztok KOSEM, Kaja DOBROVOLJC, Špela ARHAR HOLDT, Jaka ČIBEJ, Cyprian LASKOWSKI, Bojan KLEMENC, and Luka KRSNIK (2021), Frequency lists of collocations from the Gigafida 2.1 corpus, `http://hdl.handle.net/11356/1415`, Slovenian language resource repository CLARIN.SI.

Saisuresh KRISHNAKUMARAN and Xiaojin ZHU (2007), Hunting elusive metaphors using lexical resources., in *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pp. 13–20, Association for Computational Linguistics, Rochester, NY, USA, `https://aclanthology.org/W07-0103`.

George LAKOFF and Mark JOHNSON (1980), *Metaphors we live by*, University of Chicago Press, ISBN 978-0-226-46800-6.

George LAKOFF and Mark JOHNSON (2003), *Metaphors we live by*, University of Chicago Press, ISBN 0-226-46801-1.

J. Richard LANDIS and Gary G. KOCH (1977), The measurement of observer agreement for categorical data, *Biometrics*, 33(1):159–174.

Ronald W. LANGACKER (1990), *Concept, image, and symbol: The cognitive basis of grammar*, Mouton de Gruyter.

Yucheng LI, Shun WANG, Chenghua LIN, and Frank GUERIN (2023a), Metaphor detection via explicit basic meanings modelling, in Anna ROGERS, Jordan BOYD-GRABER, and Naoaki OKAZAKI, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 91–100, Association for Computational Linguistics, Toronto, Canada, doi:10.18653/v1/2023.acl-short.9, `https://aclanthology.org/2023.acl-short.9/`.

Yucheng LI, Shun WANG, Chenghua LIN, Frank GUERIN, and Loic BARRAULT (2023b), FrameBERT: Conceptual metaphor detection with frame embedding learning, in Andreas VLACHOS and Isabelle AUGENSTEIN, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1558–1563, Association for Computational Linguistics, Dubrovnik, Croatia, doi:10.18653/v1/2023.eacl-main.114, `https://aclanthology.org/2023.eacl-main.114`.

Jiahui LIANG, Stephan RAAIJMAKERS, Aletta G. DORST, and Jelena PROKIC (2024), Using large language models for conventional metaphor detection, in *Proceedings of the Workshop on Computational Approaches to Metaphor and Figurative Language*, p. 156, Bochum, Germany, `https://www.dgfs2024.ruhr-uni-bochum.de/dgfs/mam/content/dgfs2024-bochum-sprache-und-einstellung-tagungsband.pdf#page=162`.

Zhenxi LIN, Qianli MA, Jiangyue YAN, and Jieyu CHEN (2021), CATE: A contrastive pre-trained model for metaphor detection with semi-supervised learning, in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3888–3898, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, doi:10.18653/v1/2021.emnlp-main.316, `https://aclanthology.org/2021.emnlp-main.316`.

Jeannette LITTLEMORE and Graham LOW (2006), Metaphoric competence, second language learning, and communicative language ability, *Applied Linguistics*, 27(2):268–294.

Nelson F. LIU, Matt GARDNER, Yonatan BELINKOV, Matthew E. PETERS, and Noah A. SMITH (2019), Linguistic knowledge and transferability of contextual representations, in *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. 1073–1094, Association for Computational Linguistics, Minneapolis, MN, USA, doi:10.18653/v1/N19-1112, `https://aclanthology.org/N19-1112`.

Nikola LJUBEŠIĆ and Tomaž ERJAVEC (2018), Word embeddings CLARIN.SI-embed.sl 1.0, `http://hdl.handle.net/11356/1204`, Slovenian language resource repository CLARIN.SI.

Nikola LJUBEŠIĆ, Vít SUCHOMEL, Peter RUPNIK, Taja KUZMAN, and Rik VAN NOORD (2024), Language models on a diet: Cost-efficient development of encoders for closely-related languages via additional pretraining, in Maite MELERO, Sakriani SAKTI, and Claudia SORIA, editors, *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pp. 189–203, ELRA and ICCL, Torino, Italy, `https://aclanthology.org/2024.sigul-1.23/`.

Daniel LOUREIRO, Kiamehr REZAEE, Mohammad Taher PILEHVAR, and José CAMACHO-COLLADOS (2020), Language models and word sense disambiguation: An overview and analysis, *ArXiv*, abs/2008.11608, `https://api.semanticscholar.org/CorpusID:221319787`.

Rui MAO, Chenghua LIN, and Frank GUERIN (2018), Word embedding and WordNet based metaphor identification and interpretation, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1222–1231, Association for Computational Linguistics, Melbourne, Australia, doi:10.18653/v1/P18-1113, `https://aclanthology.org/P18-1113`.

Gary MASSEY (2021), Re-framing conceptual metaphor translation research in the age of neural machine translation: Investigating translators' added value with products and processes, *Training, Language and Culture*, 5(1):37–56.

Timothee MICKUS, Denis PAPERNO, Mathieu CONSTANT, and Kees VAN DEEMTER (2020), What do you mean, BERT?, in *Proceedings of the Society for Computation in Linguistics 2020*, pp. 279–290, Association for Computational Linguistics, New York, NY, USA, `https://aclanthology.org/2020.scil-1.35`.

Susan NACEY, Aletta G. DORST, Tina KRENNMAYR, and W. Gudrun REIJNIERSE, editors (2019), *Metaphor identification in multiple languages: MIPVU around the world*, John Benjamins, doi:10.1075/celcr.22.

OPENAI (2022), New and improved embedding model, `https://openai.com/index/new-and-improved-embedding-model/`, Article authors: Ryan Greene, Ted Sanders, Lilian Weng, Arvind Neelakantan.

OPENAI (2024), New embedding models and API updates, `https://openai.com/index/new-embedding-models-and-api-updates/`.

Carita PARADIS (2000), Reinforcing adjectives: A cognitive semantic perspective on grammaticalization, in Ricardo BERMÚDEZ-OTERO, David

DENISON, Richard HOGG, and C.B. MCCULLY, editors, *Generative theory and corpus studies*, pp. 233–258, Mouton de Gruyter, ISBN 3-11-016687-9.

Paolo PEDINOTTI, Eliana DI PALMA, Ludovica CERINI, and Alessandro LENCI (2021), A howling success or a working sea? Testing what BERT knows about metaphors, in Jasmijn BASTINGS, Yonatan BELINKOV, Emmanuel DUPOUX, Mario GIULIANELLI, Dieuwke HUPKES, Yuval PINTER, and Hassan SAJJAD, editors, *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 192–204, Association for Computational Linguistics, Punta Cana, Dominican Republic, doi:10.18653/v1/2021.blackboxnlp-1.13, `https://aclanthology.org/2021.blackboxnlp-1.13`.

PRAGGLEJAZ (2007), MIP: A method for identifying metaphorically used words in discourse, *Metaphor and Symbol*, 22:1–39, doi:10.1207/s15327868ms2201_1, Group authors: Peter Crisp, Raymond Gibbs, Alice Deignan, Graham Low, Gerard Steen, Lynne Cameron, Elena Semino, Joe Grady, Alan Cienki, Zoltán Kövecses.

Emily REIF, Ann YUAN, Martin WATTENBERG, Fernanda B. VIEGAS, Andy COENEN, Adam PEARCE, and Been KIM (2019), Visualizing and measuring the geometry of BERT, in H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D'ALCHÉ-BUC, E. FOX, and R. GARNETT, editors, *Advances in Neural Information Processing Systems*, volume 32, Curran Associates, Inc., `https://proceedings.neurips.cc/paper_files/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf`.

Gudrun REIJNIERSE, Christian BURGERS, Tina KRENNMAYR, and Gerard STEEN (2019), Metaphor in communication: the distribution of potentially deliberate metaphor across register and word class, *Corpora*, 14(3):301–326, doi:10.3366/cor.2019.0176.

Vassiliki RENTOUMI, George A. VOUROS, Vangelis KARKALETSIS, and Amalia MOSER (2012), Investigating metaphorical language in sentiment analysis: A sense-to-sentiment perspective, *ACM Transactions on Speech and Language Processing (TSLP)*, 9(3):1–31, doi:10.1145/2382434.2382436.

Phillip RUST, Jonas PFEIFFER, Ivan VULIĆ, Sebastian RUDER, and Iryna GUREVYCH (2021), How good is your tokenizer? On the monolingual performance of multilingual language models, in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3118–3135, Association for Computational Linguistics, Online, doi:10.18653/v1/2021.acl-long.243, `https://aclanthology.org/2021.acl-long.243`.

Pavel RYCHLÝ (2008), A lexicographer-friendly association score, in *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008*,

pp. 6–9, Masarykova Univerzita,
https://nlp.fi.muni.cz/raslan/2008/papers/13.pdf.

Shlomo SAWILOWSKY (2009), New effect size rules of thumb, *Journal of Modern Applied Statistical Methods*, 8:597–599, doi:10.22237/jmasm/1257035100.

Elena SEMINO (2008), *Metaphor in discourse*, Cambridge University Press.

Ekaterina SHUTOVA, Douwe KIELA, and Jean MAILLARD (2016), Black holes and white rabbits: Metaphor identification with visual features, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 160–170, Association for Computational Linguistics, San Diego, CA, USA, doi:10.18653/v1/N16-1020, https://aclanthology.org/N16-1020.

Ekaterina SHUTOVA, Lin SUN, and Anna KORHONEN (2010), Metaphor identification using verb and noun clustering, in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, volume 2, pp. 1002–1010, https://aclanthology.org/C10-1113/.

Gerard STEEN (2010), *A method for linguistic metaphor identification: From MIP to MIPVU*, Converging evidence in language and communication research, John Benjamins Publishing Company, ISBN 978-90-272-3903-7, doi:10.1075/celcr.14.

Gerard STEEN (2017), Deliberate metaphor theory: Basic assumptions, main tenets, urgent issues, *Intercultural Pragmatics*, 14:1–24, doi:10.1515/ip-2017-0001.

STUDENT (1908), The probable error of a mean, *Biometrika*, 6(1):1–25.

Chang SU, Shuman HUANG, and Yijiang CHEN (2017), Automatic detection and interpretation of nominal metaphor based on the theory of meaning, *Neurocomputing*, 219:300–311, doi:10.1016/j.neucom.2016.09.030.

Karen SULLIVAN (2013), *Frames and constructions in metaphoric language*, John Benjamins, doi:10.1075/cal.14.

Paul THIBODEAU and Lera BORODITSKY (2011), Metaphors we think with: The role of metaphor in reasoning, *PloS One*, 6:e16782, doi:10.1371/journal.pone.0016782.

Peter TURNEY, Yair NEUMAN, Dan ASSAF, and Yohai COHEN (2011), Literal and metaphorical sense identification through concrete and abstract context, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 680–690, Association for Computational Linguistics, Edinburgh, Scotland, https://aclanthology.org/D11-1063.

Matej ULČAR and Marko ROBNIK-ŠIKONJA (2020), High quality ELMo embeddings for seven less-resourced languages, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4731–4738, European Language Resources Association, Marseille, France, ISBN 979-10-95546-34-4, https://aclanthology.org/2020.lrec-1.582.

Matej Ulčar and Marko Robnik-Šikonja (2021), Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0, http://hdl.handle.net/11356/1397, Slovenian language resource repository CLARIN.SI.

Matej Ulčar and Marko Robnik-Šikonja (2023), Sequence-to-sequence pretraining for a less-resourced Slovenian language, *Frontiers in Artificial Intelligence*, 6:932519, doi:10.3389/frai.2023.932519.

Matej Ulčar, Aleš Žagar, Carlos S. Armendariz, Andraž Repar, Senja Pollak, Matthew Purver, and Marko Robnik-Šikonja (2021), Evaluation of contextual embeddings on less-resourced languages, *ArXiv*, doi:10.48550/arXiv.2107.10614, version 1.

Matej Ulčar and Marko Robnik-Šikonja (2020), FinEst BERT and CroSloEngual BERT: less is more in multilingual models, in P. Sojka, I. Kopeček, K. Pala, and A. Horák, editors, *Text, Speech, and Dialogue TSD 2020*, volume 12284 of *Lecture Notes in Computer Science*, Springer, doi:10.1007/978-3-030-58323-1_11.

Akira Utsumi (2011), Computational exploration of metaphor comprehension processes using a semantic space model, *Cognitive Science*, 35(2):251–296, doi:10.1111/j.1551-6709.2010.01144.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1551-6709.2010.01144.x.

Ada Vidovič-Muha (1978), Merila pomenske delitve nezaimenske pridevniške besede [Criteria for the semantic classification of non-pronominal adjectival words], *Slavistična revija*, 26(3):253–276, https://srl.si/ojs/srl/article/view/COBISS_ID-21953378.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov (2019), Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Association for Computational Linguistics, Florence, Italy, doi:10.18653/v1/P19-1580, https://aclanthology.org/P19-1580.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen (2020), Probing pretrained language models for lexical semantics, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7222–7240, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.emnlp-main.586, https://aclanthology.org/2020.emnlp-main.586.

Lennart Wachowiak and Dagmar Gromann (2023), Does GPT-3 grasp metaphors? Identifying metaphor mappings with generative language models, in Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1018–1032, Association for

Computational Linguistics, Toronto, Canada,
doi:10.18653/v1/2023.acl-long.58,
https://aclanthology.org/2023.acl-long.58.

Shun WANG, Yucheng LI, Chenghua LIN, Loic BARRAULT, and Frank GUERIN (2023), Metaphor detection with effective context denoising, in Andreas VLACHOS and Isabelle AUGENSTEIN, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1404–1409, Association for Computational Linguistics, Dubrovnik, Croatia, doi:10.18653/v1/2023.eacl-main.102,
https://aclanthology.org/2023.eacl-main.102.

Shun WANG, Ge ZHANG, Han WU, Tyler LOAKMAN, Wenhao HUANG, and Chenghua LIN (2024), MMTE: Corpus and metrics for evaluating machine translation quality of metaphorical language, in Yaser AL-ONAIZAN, Mohit BANSAL, and Yun-Nung CHEN, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11343–11358, Association for Computational Linguistics, Miami, FL, USA,
doi:10.18653/v1/2024.emnlp-main.634,
https://aclanthology.org/2024.emnlp-main.634/.

Yile WANG and Yue ZHANG (2024), Lost in context? On the sense-wise variance of contextualized word embeddings, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:639–650, doi:10.1109/TASLP.2023.3337643.

Bernard Lewis WELCH (1947), The generalization of 'Student's' problem when several different population variances are involved, *Biometrika*, 34(1–2):28–35, doi:10.1093/biomet/34.1-2.28.

Gregor WIEDEMANN, Steffen REMUS, Avi CHAWLA, and Chris BIEMANN (2019), Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings, *ArXiv*, abs/1909.10430.

Yorick WILKS (1978), Making preferences more active, *Artificial Intelligence*, 11(3):197–223, doi:10.1016/0004-3702(78)90001-2, https://www.sciencedirect.com/science/article/pii/0004370278900012.

Linting XUE, Noah CONSTANT, Adam ROBERTS, Mihir KALE, Rami AL-RFOU, Aditya SIDDHANT, Aditya BARUA, and Colin RAFFEL (2021), mT5: A massively multilingual pre-trained text-to-text transformer, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Association for Computational Linguistics, Online, doi:10.18653/v1/2021.naacl-main.41,
https://aclanthology.org/2021.naacl-main.41.

Omnia ZAYED, John Philip MCCRAE, and Paul BUITELAAR (2018), Phrase-level metaphor identification using distributed representations of word meaning, in *Proceedings of the Workshop on Figurative Language Processing*, pp. 81–90, Association for Computational Linguistics, New Orleans, LA, USA,
doi:10.18653/v1/W18-0910, https://aclanthology.org/W18-0910.

Kaitlyn ZHOU, Kawin ETHAYARAJH, Dallas CARD, and Dan JURAFSKY (2022), Problems with cosine as a measure of embedding similarity for high frequency words, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 401–423, Association for Computational Linguistics, Dublin, Ireland, doi:10.18653/v1/2022.acl-short.45, https://aclanthology.org/2022.acl-short.45.

Ana ZWITTER VITEZ, Mojca BRGLEZ, Marko ROBNIK ŠIKONJA, Tadej ŠKVORC, Andreja VEZOVNIK, and Senja POLLAK (2022), Extracting and analysing metaphors in migration media discourse: towards a metaphor annotation scheme, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2430–2439, European Language Resources Association, Marseille, France, https://aclanthology.org/2022.lrec-1.259.

Mojca Brglez
0000-0002-8806-0942

Faculty of Arts
University of Ljubljana
Aškerčeva 2, 1000 Ljubljana, Slovenia

“Jožef Stefan” Institute
Jamova 39, 1000 Ljubljana, Slovenia

Špela Vintar
0000-0003-1934-0200

Faculty of Arts
University of Ljubljana
Aškerčeva 2, 1000 Ljubljana, Slovenia

“Jožef Stefan” Institute
Jamova 39, 1000 Ljubljana, Slovenia

Mojca Brglez and Špela Vintar (2025), In search of semantic distance: metaphorical and non-metaphorical constructions in static and contextual embeddings, *Journal of Language Modelling*, 13(2):207–260
https://dx.doi.org/10.15398/jlm.v13i2.437