Montague semantics and modifier consistency measurement in neural language models

Danilo S. Carvalho[1], **Edoardo Manino**[2], **Julia Rozanova**[2],
Lucas Cordeiro[2] and **André Freitas**[1,2,3]
National Biomarker Centre, CRUK-MI, Univ. of Manchester, United Kingdom[1]
Department of Computer Science, University of Manchester, United Kingdom[2]
Idiap Research Institute, Switzerland[3]
{firstname.lastname}@manchester.ac.uk

Abstract

This work proposes a novel methodology for measuring compositional behavior in contemporary language embedding models. Specifically, we focus on adjectival modifier phenomena in adjective-noun phrases. In recent years, distributional language representation models have demonstrated great practical success. At the same time, the need for interpretability has elicited questions on their intrinsic properties and capabilities. Crucially, distributional models are often inconsistent when dealing with compositional phenomena in natural language, which has significant implications for their *safety* and *fairness*. Despite this, most current research on compositionality is directed towards improving their performance on similarity tasks only. This work takes a different approach, introducing three novel tests of compositional behavior inspired by Montague semantics. Our experimental results indicate that current neural language models do not behave according to the expected linguistic theories. This indicates that current language models may lack the capability to capture the semantic properties we evaluated on limited context, or that linguistic theories from Montagovian tradition may not match the expected capabilities of distributional models.

1 Introduction

Distributional semantics and neural language models have been a dominant approach in language representation models for nearly a decade since the emergence of deep learning methods (Lenci et al., 2022). This is due to the consistent achievements in terms of the state-of-the-art performance in various downstream NLP tasks and the progressive increase of their parameter size and complexity. Interest in the properties of these models and their relationships with semantic formalisms is older than their rise to mainstream use (Baroni and Zamparelli, 2010). However, the recent demand for models delivering safety guarantees and better inference control has highlighted its importance (Floridi and Chiriatti, 2020). This is of particular relevance to retrieval-augmented generation (RAG) (Lewis et al., 2020), as implicit compositional assumptions are a source of semantic gaps in natural language queries.

Indeed, understanding the intrinsic linguistic and semantic properties of distributional neural language models can provide important insight on their capabilities and limitations. From a purely distributional perspective, studies have been conducted on analysing the concept drift (Sommerauer and Fokkens, 2019) and biases (Bhardwaj et al., 2021) of such models. On a linguistic front, attempts at mapping vector representations to dictionary senses and lexical features have yielded promising results (Pilehvar and Navigli, 2015; Carvalho and Nguyen, 2017). Similarly, works that probed for the presence of linguistic features in sentence-level representations revealed a wide array of syntactic information captured (Miaschi and Dell'Orletta, 2020; Ferreira et al., 2021).

However, one issue that has been underexplored from a linguistic standpoint is compositionality and their associated set-theoretic (Montagovian) concepts, where efforts have been directed towards improving performance of the representations on similarity tasks (see Section 5), without attempting to relate the linguistic principles involved with compositional properties observed.

This work proposes to fill this research gap, electing the *modifier phenomena* (Dixon et al., 2004; Morzycki, 2016) as a starting point for the analysis of compositional properties in language models, and adopting text embeddings as proxies of concept denotations. In this way, we can test the manifestation of compositional properties in adjective phrase denotations, such as intersectivity, as a function of the consistency of geometric properties in the embedding space, in the form of *metamor-*

Proceedings of the 31st International Conference on Computational Linguistics, pages 5515–5529
January 19–24, 2025. ©2025 Association for Computational Linguistics

phic relations (Chen et al., 2018). The hypothesis of proxying denotations through embeddings has been implicitly used for the "vector analogy" tasks (e.g., $king - man + woman = queen$) (Mikolov et al., 2013b), but is used here explicitly to test denotation properties in the embedding space.

Similarly, the concept of metamorphic relation has recently waded its way from the field of software engineering (Chen et al., 2018) to machine learning and natural language processing (Belinkov and Bisk, 2018; Manino et al., 2022). There, it brings the promise of formally defining the expected behavior of a learning-based model and rigorously testing whether it holds in practice without the need for ground-truth labels. Popular applications of behavioral testing usually focus on plain substitutions of similar words (e.g., robustness to synonym replacement) (Jia et al., 2019), or semantic opposition (e.g., changing the gender of nouns) (Ma et al., 2020). However, efforts have been made to extend this framework to higher-level linguistic properties such as systematicity and transitivity (Manino et al., 2022). The present work continues this line of research by grounding the concept of metamorphic relation onto the linguistic tradition of formal semantics.

Hypothesis [embedding-denotation analogy]: Assume that the modifier phenomena is described by a Set representation/Montague semantics compositional model. We expect a large language model, which at the limit captures the distributional properties of an infinite corpus of utterances, to show empirical evidence of the formal properties of the modifier phenomena.

Research Questions: In this paper, we restrict our inquiry to adjective modifiers and contemporary neural language models. In this setting, we can pose the following research questions:

RQ1. Adjective-noun composition is described in Montague semantics as a function mapping elements between two sets A $\rightarrow$ P corresponding to the properties satisfied by the individuals referred by each set (denotation). Can we expect to observe a correspondence of these theoretical linguistic properties in neural language models that operate on dense vector spaces?

RQ2. Existing neural language models are limited by their choices of the learning process (objective functions) and the language data available for model training. To what degree can we observe evidence of the compositional effect of adjective modifiers? Do contextual models differ from non-contextual ones in this regard?

Contributions: We propose a methodology for measuring the presence of compositional behaviour in contemporary neural language models related to adjectival modifier phenomena in adjective-noun phrases, from a Montagovian formalism perspective. Our methodology *translates a set-based formal semantic theory into metamorphic relations in embedding spaces* based on the cosine distance between embeddings (RQ1). Our results show that current neural language models do not behave consistently according to the linguistic theories with regard to the evaluated intersective property. In fact, there is no statistically significant difference between different adjective categories: the empirical behaviour we observe tends to be *intersective* across all inputs and language models (RQ2). Additionally, we found that while large SOTA transformer models behave similarly to non-contextual models regarding intersectivity, when accounting for mean-pooling bias, they largely differ in terms of subsectivity, placing heavy emphasis on adjectives instead of nouns (RQ2). The results indicate that current language models may lack the capability to capture the semantic properties we evaluated on limited context, or that linguistic theories from Montagovian tradition may not match expected capabilities of distributional models. Finally, we make publicly available the developed experimental pipeline and dataset (44652 adjective-noun phrases) for reproducibility purposes[1].

Scope of the study: The formal linguistic properties evaluated in this study are not sufficient nor intended to evaluate general compositionality, but are a fundamental part of a larger set of compositional phenomena, which includes verbal, nominal and even non-linguistic (e.g., arithmetic) composition. Additionally, the proposed methodology focuses on the model's distributional embedding spaces, rather than specific downstream tasks. This allows it to be applied for general assessment of current and future models' compositional behaviour, irrespective of task capability or specialisation.

The remainder of this paper is organized as follows: Section 2 explains the linguistic grounding of this work in more detail, Section 3 discusses our methodology, Section 4 reports our experimental setup and discusses its findings, Section 5 presents the broader landscape of related works, and finally

[1] Experimental code and the full dataset are available at: https://github.com/dscarvalho/modifiers_consistency

Section 6 summarizes our contribution and concludes with some final remarks.

2 The modifier phenomena

Of all linguistic phenomena arising from the composition of meaning of two or more words, *modification*, and in particular the application of adjectives, has been the subject of extensive study (Dixon et al., 2004; Morzycki, 2016).

2.1 Modification semantics

From a linguistic standpoint, *modification* does not constitute a single grammatical phenomenon, being a term for expressions that do not fit into either the predicate or argument categories. In fact, *modification* characterizes both a family of (internal) lexical semantic characteristics and of (external) distributional ones (Morzycki, 2016). For the purpose of our study, we narrow down the definition of modifiers to a set of compositional principles regarding intensional interpretations from a Montagovian formalism, with adjective phrases being the object of analysis (Boleda et al., 2013; Paperno and Baroni, 2016). On the one hand, this choice allows us to interpret nouns as their denotations in set theoretical form. For example, we assume that the noun "dog" represents the set of properties that hold for any individual to which the concept of dog applies. On the other hand, this choice allows us to interpret adjectives as functions of set-based denotations. The latter is discussed below.

2.2 Adjective types and interpretations

Adjectives can be classified according to their effect on the denotations they modify (Morzycki, 2016; Pavlick and Callison-Burch, 2016):

- **Intersective** (or *extensional*): describes the intersection of the noun denotation with the denotation of the adjective itself. Thus, the adjective can also be interpreted as a set. E.g. "red car" denotes the set of things that are both "car" and "red".

- **Subsective (non-intersective)**: describes a strict subset of the noun denotation it modifies. E.g. "skillful teacher" denotes a subset of teachers, but there is no general denotation for "skillful".

- **Privative non-subsective**: describes a set that is completely disjoint from the denotation of the noun it modifies. E.g. "fake wall" denotes a set of things that are definitely *not* walls.

- **Plain non-subsective**: describes a set that may or may not be a subset of the noun denotation it modifies, depending on the context or the adjective itself. E.g. "alleged criminal" denotes a set of individuals whose inclusion in the set of criminals is dubious or undefined, while "former president" denotes a set of individuals that are not presidents anymore.

- **Ambiguous**: can be applied to any of the previous categories, depending on the context and the modified noun. E.g. "big" is intersective in the phrase "big truck" and subsective non-intersective in the phrase "big fool".

Section 3.1 contains further formalization of these adjective types and their related properties.

2.3 Distributional questions

Hanging fundamentally on the *distributional hypothesis*, distributional models are primarily optimised for capturing statistical co-occurrence relations (syntagmatic and paradigmatic relations) at scale. As a result, distributional models naturally excel at computing measures of semantic *relatedness* and semantic *similarity* between any given pair of terms in a corpus. However, their ability at capturing more structured compositional behaviour is unclear.

Efforts at building distributional models that exhibit compositional behaviour *by construction* has been made in the past (Clark and Pulman, 2007; Mitchell and Lapata, 2008; Guevara, 2010). Unfortunately, these efforts predate the advent of state-of-the-art self-supervised language models, and cannot compete with their performance. In fact, recent language models have tackled composition in more implicit ways, with state-of-the-art approaches being trained on multiple objectives such as masked word prediction, sentence-level similarity and entailment functions (Reimers and Gurevych, 2019; Sanh et al., 2019; Ni et al., 2022).

This raises the question of whether the representations obtained in this way could be employed as *proxies* for word and phrase denotations. If this is the case, then any term comparisons made in the embedding spaces would represent an equivalent operation between denotations (e.g., subset inclusion). Conversely, set theoretical properties on denotations could be interpreted as geometrical

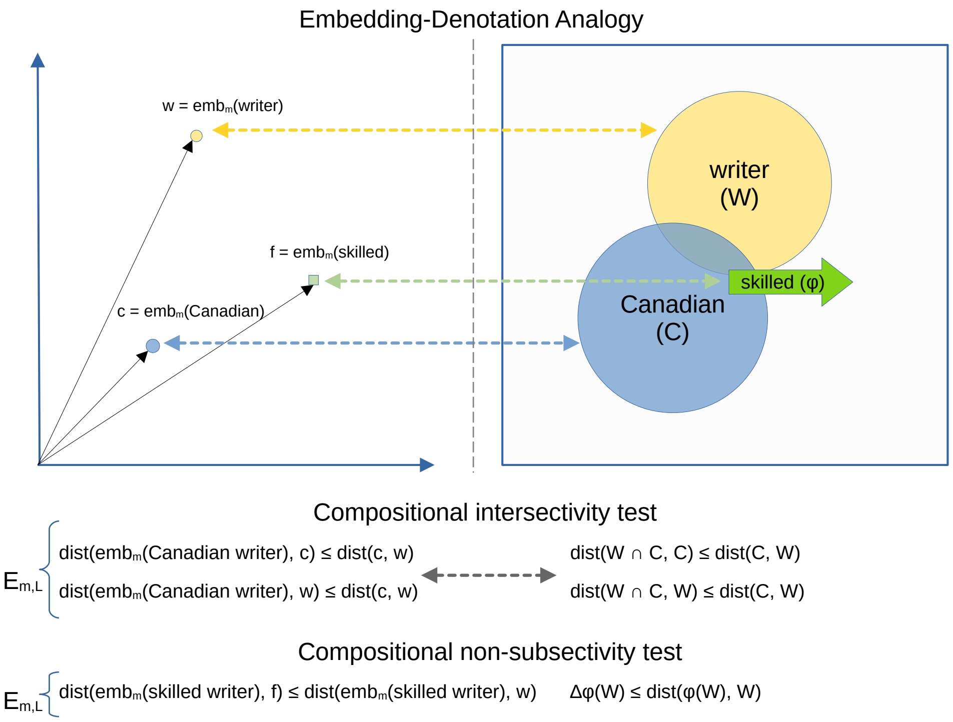
Embedding-Denotation Analogy
w = emb_m(writer)
f = emb_m(skilled)
c = emb_m(Canadian)
writer (W)
skilled (φ)
Canadian (C)
Compositional intersectivity test
E_m,L
dist(emb_m(Canadian writer), c) ≤ dist(c, w)
dist(emb_m(Canadian writer), w) ≤ dist(c, w)
dist(W ∩ C, C) ≤ dist(C, W)
dist(W ∩ C, W) ≤ dist(C, W)
Compositional non-subsectivity test
E_m,L
dist(emb_m(skilled writer), f) ≤ dist(emb_m(skilled writer), w)
Δφ(W) ≤ dist(φ(W), W)

Figure 1: Methodology for testing model consistency regarding the modifier phenomena in adjective-noun phrases (ANs). m represents a language model and L the regular language $(adj\ )+noun$. $E_{m,L}$ is calculated by averaging the no. of combinations $(a, n, \phi, p = an)$ where the inequalities hold over the vocabulary size.

properties of the embedding space (e.g., vector distance constraints). This understanding lies at the foundation of the methodology presented hereon.

3 Methodology

Our methodology is centred around the hypothesis that neural embeddings should correctly approximate the linguistic denotation of the input phrases. In this light, we propose three different metamorphic tests to check whether neural models satisfy such hypothesis.

3.1 Set-based phrase denotations

In general, we say that a noun n can be modified by an adjective a to form an adjective-noun phrase $p = an$. The denotation of p can be represented as a set, and depends on the type of the adjective a (see Section 2). More specifically, we divide the adjectives into two main categories: *intersective* and *non-intersective*. Here, the *non-intersective* category includes both subsective and non-subsective adjectives.

On the one hand, if a is an intersective adjective, then the denotation of p is simply the intersection of the denotations of a and n. For example, the intersective phrase $p = Canadian\,writer$ is associated with the following Montague denotations (intensions):

$$\begin{aligned} n(x) &= \lambda x.[writer(x)] \\ a(x) &= \lambda x.[Canadian(x)] \\ p(x) &= \lambda x.[a(x) \wedge n(x)] \end{aligned}$$

and corresponding sets (extensions):

$$\begin{aligned} N &\equiv \{x \mid n(x) = \top\} \\ A &\equiv \{x \mid a(x) = \top\} \\ P &\equiv A \cap N \end{aligned} \tag{1}$$

where $P \subseteq N$ and $P \subseteq A$.

On the other hand, if a is a non-intersective adjective, then the denotation of p involves functions over sets. For example, the phrase $p = skilled\,writer$ requires the following Montague denotations:

$$\begin{aligned} a(n, x) &= \lambda n.\lambda x[skilled(n(x), x)] \\ p(x) &= \lambda x.[a(W, x)] \end{aligned}$$

where function a can discriminate whether x is a skilled writer, but has no concept of "skilfulness" in general. Accordingly, the corresponding sets (extensions) are:

$$P \equiv A \equiv \{x \mid p(x) = \top\} \subseteq N \tag{2}$$

Note that in the intersective case (see Equation 1) the set P is included in both A and N, whereas in the non-intersective case (see Equation 2) this is not the case. As a result, if we could measure the distance between these three sets for a generic adjective-noun phrase $p = an$, then we should be able to identify the type of the adjective a. Figure 2 illustrates this concept of relations between sets.

3.2 Embedding-denotation analogy

Thus, our core hypothesis is the following. If the phrase embedding correctly represents its denotation, we should observe some analogous inclusion relations between them. Since embeddings are defined in vector space, the inclusion relations must be replaced with another appropriate measure (e.g., cosine, Euclidean). This hypothesis motivates the following tests.

3.3 Testing intersectivity (single phrase)

Assume $p = t_1 t_2 \dots t_h$ is an adjective-noun phrase containing one or more adjectives. If all adjectives were intersective, the corresponding set relations $P \subseteq T_i$ would be satisfied (see Section 3.1). In contrast, any two individual terms t_j, t_k are generally unrelated, yielding $T_j \nsubseteq T_k$. We hypothesise that set inclusion translates into shorter distances between embedding, which leads us to the following test of intersectivity:

$$\begin{aligned} I_{m,p} \equiv\, & d(emb_m(p), emb_m(t_i)) \\ & \leq d(emb_m(t_j), emb_m(t_k)) \\ & \forall i, j, k; \;\; j < k \end{aligned} \tag{3}$$

where $t_{1..h}$ is a term of the phrase p and emb_m is the embedding function for model m. We define the consistency of a model m concerning Equation 3 by taking the expectation of its truth value:

$$E_{m,L}\{I_{m,p} = \top\}, \quad p \sim L \tag{4}$$

where L is the regular language "$(adj\ \ ) + noun$" with alphabet Σ, and p is extracted from L according to a probability distribution.

3.4 Testing intersectivity (phrase pairs)

Here, we give a more complex distance relation between pairs of adjective-noun phrases, which allows us to test for behaviour that result strictly from intersective effects, by controlling for the induced intersectivity from vector pooling while accounting for synonymy, if present. Define $p_{a_1 n_1}$, $p_{a_1 n_2}$, $p_{a_2 n_1}$ and $p_{a_2 n_2}$ as all possible concatenations of two adjectives $a_1 \neq a_2$ and two nouns $n_1 \neq n_2$. We expect the following metamorphic relation to holding:

$$\begin{aligned} II_{m,\{p\}} = \,& d(emb_m(p_{a_1 n_1}), emb_m(p_{a_1 n_2})) \\ & \leq d(emb_m(p_{a_2 n_1}), emb_m(p_{a_2 n_2})) \end{aligned} \tag{5}$$

when a_1 is intersective and a_2 is not. For example:

$$\begin{aligned} & d(Canadian\, writer, Canadian\, surgeon) \\ & \leq d(skillful\, writer, skillful\, surgeon) \end{aligned}$$

where $Canadian$ is an intersective adjective and $skilful$ is not. This is because we expect a $Canadian\, writer$ to have something in common with a $Canadian\, surgeon$, i.e., the fact that they are both Canadian. In contrast, a $skilful\, writer$ and a $skilful\, surgeon$ are not similar as there is minimal overlap between their skills.

As for Equation 4, we call the consistency of m:

$$E_{m,L^2}\{II_{m,\{p\}} = \top\}, \quad \{p\} \sim L^2 \tag{6}$$

where $\{p\} \equiv \{a_1 n_1, a_1 n_2, a_2 n_1, a_2 n_2\}$. The value of Equation 6 should approach 1.0 when all a_1 in L^2 are intersective and all a_2 are not.

3.5 Testing non-subsectivity

We propose to test for non-subsectivity by looking at the relative change caused by an adjective to a noun when combined in a phrase. Let $p = an$ be an adjective-noun phrase with associated set P and noun set N. Subsective composition guarantees $P \subseteq N$, whereas non-subsective composition does not. Consequently, we hypothesise that the embedding of p is closer to n when a is subsective. Accordingly, we can test for non-subsectivity with the following metamorphic relation:

$$\begin{aligned} NI_{m,p} = \,& d(emb_m(p), emb_m(a)) \\ & \leq d(emb_m(p), emb_m(n)) \end{aligned} \tag{7}$$

and the corresponding consistency metric:

$$E_{m,L}\{NI_{m,p} = \top\}, \quad p \sim L \tag{8}$$

where L is the same language as in Equation 4.

4 Experimentation and discussion

4.1 Experimental setup

To perform the tests introduced in Section 3, we need the following components: a measure of distance d in embedding space, a set of adjective-noun phrases covering all adjective types (the input data), and the language models to be tested. In all our experiments we use the cosine distance, the input phrases and the language models described below.

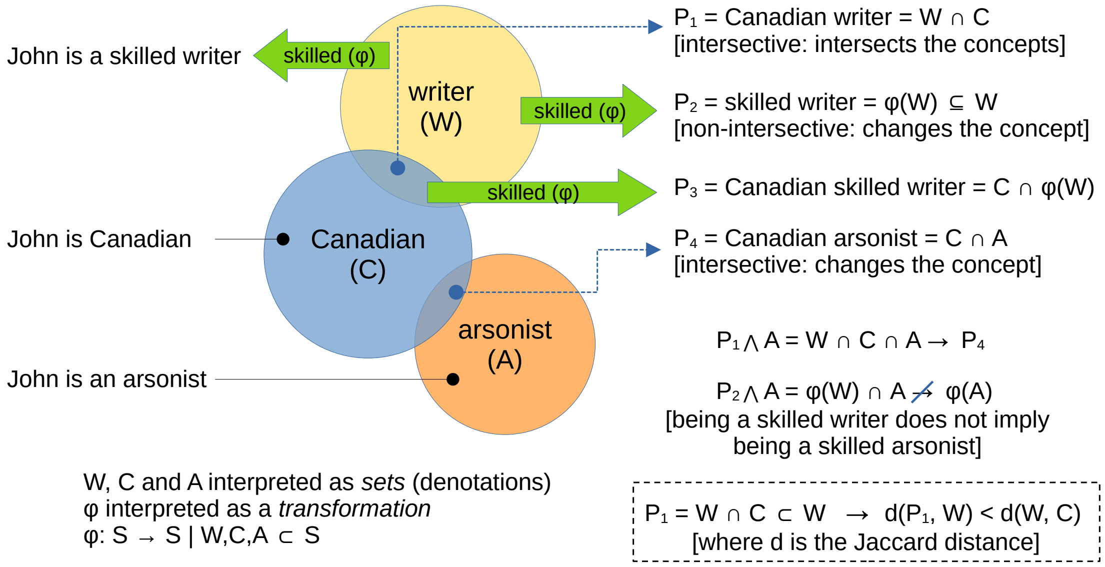
John is a skilled writer
skilled (φ)
writer (W)
skilled (φ)
skilled (φ)
John is Canadian
Canadian (C)
John is an arsonist
arsonist (A)
P₁ = Canadian writer = W ∩ C
[intersective: intersects the concepts]
P₂ = skilled writer = φ(W) ⊆ W
[non-intersective: changes the concept]
P₃ = Canadian skilled writer = C ∩ φ(W)
P₄ = Canadian arsonist = C ∩ A
[intersective: changes the concept]
P₁ ∧ A = W ∩ C ∩ A → P₄
P₂ ∧ A = φ(W) ∩ A ↛ φ(A)
[being a skilled writer does not imply being a skilled arsonist]
W, C and A interpreted as sets (denotations)
φ interpreted as a transformation
φ: S → S | W,C,A ⊂ S
P₁ = W ∩ C ⊂ W → d(P₁, W) < d(W, C)
[where d is the Jaccard distance]

Figure 2: Intersective and non-intersective set relations in adjective-noun phrase denotations.

4.1.1 Data collection

We consider adjective categories based on Morzycki (2016) and Pavlick and Callison-Burch (2016), where the latter provides a further subdivision of non-subsective adjectives. We select the examples in (Morzycki, 2016) for the sets of subsective adjectives, and we use the dataset in (Pavlick and Callison-Burch, 2016) for the collection of non-subsective adjectives, both summing 61 adjectives. For each adjective in this initial list, a closest synonym was chosen for the phrase pair intersectivity test, totaling 122 adjectives. At the same time, we choose a set of 12 nouns covering both concrete and abstract concepts to form adjective/noun phrases.

The adjective and nouns lists were reviewed by each of the authors. While most adjective categorisations were left unchanged, we included an “ambiguous” category to house those adjectives that had ambiguous meaning within our phrase set. The categories and their definitions are presented in Table 1. The complete list of categorised adjectives is included as supplementary material (Appendix A).

4.1.2 Phrase Generation

The phrases were generated by using a regular language defined by the expression $(adj\ )+noun$, where adj and $noun$ are taken from the lists of adjectives and nouns respectively. More formally, $adj = (wild|red|...)$ and $noun = (student|dog|...)$. All the phrases up to 3 words were generated: e.g. “wild dog” and “square assumed law”. The final dataset contains *44652 phrases*[1].

For reasons of space, we introduce a shorthand notation for the two types of phrases we generate: we write AN (respectively, AAN) to denote a phrase composed of a single adjective followed by a noun (respectively, two adjectives followed by a noun). With slight abuse of notation, we also use AN and AAN to refer to the set interpretation (denotation) of a phrase rather than the phrase itself.

4.1.3 Encoding Strategy & Language Models

As we investigate emergent compositional behaviour, we selected models that provide a single sentence representation rather than a sequence of token representations. Most often than not, these are variants of state-of-the-art transformer-based model. However, they are further trained to generate composed vector representations which are more informative than, for example, mean pooling of token representations. More specifically, we consider DPR (Karpukhin et al., 2020), LaBSE (Feng et al., 2022), Specter (Cohan et al., 2020), OpenAI’s text-embeddings-3-small [TE3-small] (OpenAI, 2024), NV-Embed-v2 (Lee et al., 2024) and Stella[en_1.5B_v5] ([@HuggingFace], 2024). The last two being respectively the current first and third ranked at the MTEB benchmark (Muennighoff et al., 2023). With the exception of TE3-small, which is closed-source, all selected transformer models compose token representations either by CLS hidden state pooling (DPR, LaBSE, Specter) or by a specialised attention model (NV-Embed-v2,

Adjective Type	Set-Theoretic Definition	Examples	# of Adjectives
Subsective (Intersective)	$AN \subseteq N$ and $AN \subseteq A$	Red, Wild	22
Subsective (Non-Intersective)	$AN \subseteq N$ and $AN \nsubseteq A$	Skilful, Rare	12
Non-Subsective (Plain)	$AN \nsubseteq N$ and $AN \cap N \neq \emptyset$	Alleged, Disputed	54
Non-Subsective (Privative)	$AN \cap N = \emptyset$	Fake, Imaginary	28
Ambiguous	Contextually, one of the above	Old, Big	6

Table 1: Adjective type composition for the vocabulary.

Models	Adjective Type				
	S-I	S-NI	NS-Pl	NS-Pr	A
DPR	0.86	0.90	0.85	0.89	0.97
LaBSE	1.0	1.0	1.0	1.0	1.0
Specter	0.93	0.99	0.97	0.93	0.97
TE3-small	1.0	1.0	1.0	1.0	1.0
NV-Embed-v2	0.73	0.67	0.8	0.85	0.75
stella_en_1.5B_v5	1.0	1.0	1.0	1.0	1.0
Glove	1.0	1.0	1.0	1.0	1.0
Word2Vec	1.0	1.0	1.0	1.0	1.0

Table 2: Consistency scores of the intersective property in Equation 4, for single adjective-noun phrases (AN format). We use the following shorthand notation in the columns: Ambiguous (A), Subsective-Intersective (S-I), Subsective Non-Intersective (S-NI), Plain Non-Subsective (NS-Pl), Privative Non-Subsective (NS-Pr).

Models	Adjective Type Pair				
	(S-I, S-I)	(S-NI, S-I)	(NS-Pl, S-I)	(NS-Pr, S-I)	(A, S-I)
DPR	0.52	0.43	0.53	0.52	0.62
LaBSE	0.92	0.93	0.95	0.91	0.97
Specter	0.67	0.73	0.72	0.67	0.73
TE3-small	1.0	1.0	1.0	1.0	1.0
NV-Embed-v2	0.78	0.71	0.68	0.81	0.75
stella_en_1.5B_v5	1.0	1.0	1.0	1.0	1.0
Glove	1.0	1.0	1.0	0.94	1.0
Word2Vec	1.0	1.0	0.97	0.94	1.0

Table 3: Consistency scores of the intersective property in Equation 4, for adjective-noun phrases with two adjectives (AAN format). Same notation as Table 2. Results for all type combinations are included as supplementary material (Appendix B).

Stella). This avoids the inherent intersective bias from mean pooling.

For comparison, we also run the experiments on non-contextual language models trained on a purely distributional objective. In particular, we use mean-pooled representations of the Word2Vec (Mikolov et al., 2013a) and Glove (Pennington et al., 2014) models. These models provide a useful baseline to compare the aforementioned contextual models against.

4.2 Results and discussion

4.2.1 Intersectivity experiment (single phrase)

Our first metamorphic property from Section 3.3 requires that the embedding of an adjective-noun phrase lies closer to each term than the distance between any pair of terms. The results in Table 2 indicate that except for DPR, Specter and NV-Embed-v2, the models' pooling operations are equivalent to mean pooling, making them universally intersective. If we interpret the metric in Equation 4 as indicative of intersective behaviour, we would have to conclude that the remaining models do not behave according to the expected modifier phenomena formalisms, with higher consistency scores on non-intersective pairings. This conclusion is corroborated when we consider the results on phrases with two adjectives (AAN) in Table 3, which further highlights the differences between the models' compositional properties. Those can be summarised in the following findings:

Models with mean-pooling equivalent composition are universally intersective (vice-versa)
As averaging embeddings will always produce one that is the closest to both, satisfying Equation 3. Conversely, a universally intersective model is likely to have mean-pooling equivalent composition. This can be observed on LaBSE, TE3-small and Stella, which are not mean-pooling models.

Models without mean-pooling equivalent composition do not consistently capture adjective intersectivity
As observed on DPR, Specter and NV-Embed-v2, the embedding distance relations are dependent on attention parameters not corresponding with the adjective categorisation.
More details are available on Appendix B.

Models	Adjective Type Pair				
	(S-I, S-I)	(S-I, S-NI)	(S-I, NS-Pl)	(S-I, NS-Pr)	(S-I, A)
DPR	0.50	0.32	0.34	0.50	0.42
LaBSE	0.50	0.42	0.34	0.53	0.33
Specter	0.50	0.65	0.55	0.50	0.57
TE3-small	0.50	0.51	0.48	0.48	0.82
NV-Embed-v2	0.50	0.54	0.51	0.51	0.82
stella_en_1.5B_v5	0.50	0.75	0.64	0.58	0.91
Glove	0.50	0.66	0.69	0.70	0.47
Word2Vec	0.50	0.75	0.65	0.49	1.0

Table 4: Consistency score of the intersective property in Equation 6, for pairs of adjective-noun phrases with a single adjective (AN format). Same notation as Table 2.

Models	Adjective Type				
	S-I	S-NI	NS-Pl	NS-Pr	A
DPR	0.46	0.37	0.48	0.54	0.39
LaBSE	0.36	0.31	0.51	0.33	0.19
Specter	0.48	0.31	0.49	0.57	0.33
TE3-small	0.81	0.75	0.74	0.77	0.39
NV-Embed-v2	0.84	0.79	0.79	0.83	0.81
stella_en_1.5B_v5	0.81	0.56	0.58	0.64	0.33
Glove	0.61	0.22	0.22	0.32	0.28
Word2Vec	0.55	0.21	0.34	0.49	0.0

Table 5: Consistency score (consistency) of the non-subsective property in Equation 8, for single adjective-noun phrases (AN format). Same notation as Table 2.

4.2.2 Intersectivity experiment (phrase pair)

Our second metamorphic property from Section 3.4 completes the picture on intersectivity. The property requires adjective-noun phrases that share the same intersective adjective to be closer to each other than phrases with non-intersective ones. Table 4 reports the results of our experiments, which suggest that each model places intersective emphasis in a different category of adjectives, with Stella being the one that most closely approaches the linguistically expected behaviour, together with the non-contextual baselines.

4.2.3 Non-subsectivity experiment

Our third metamorphic relation from Section 3.5 requires the adjective to "pull" the embedding of the whole phrase closer to them than the associated noun. This is a reasonable requirement because non-subsective adjectives completely change the meaning of the noun, rather than just specializing it. Our final experiment allows us to test whether this is indeed the behaviour of contemporary language models.

The results are in in Table 5. Here, there is a clear trend regarding the size of the models, with the larger ones (in order of size: Stella, NV-Embed-v2, TE3-small) having substantially higher consistency scores overall. This indicates that those models place a much larger weight on the adjectives than the nouns in the composition process. The key findings of this experiment can be summarised as follows:

None of the tested models behave according to the expectations given by the subsectivity formalism
The consistency scores show different patterns of subsectivity w.r.t. the categories for each model, but in none of them the highest score belongs to a 'NS' category.

Larger models composition process largely emphasises adjectives instead of nouns
This effect is mostly independent of the adjective category, with the exception of ambiguous ones, which can be observed in TE3-small and Stella.

A case of particular interest is the ambiguously typed adjectives (dependent on the represented word sense): we see that the models do not always seem to agree on the chosen sense. The numerical behaviour hints at whether the model is more likely to choose intersective or non-intersective senses of adjectives such as "old".

Thus, adjective type differences display relatively low compositional effects on broad intersectivity and subsectivity in the evaluated models. This phenomenon indicates that while such differences may be encoded in individual word representations, specially on non-contextual models, they do not transfer generally in the expected way to the compositions.

5 Related work

Before the advent of self-supervised language models, much work has gone into constructing formally-motivated vector representations. To this end, Clark and Pulman (2007) employs tensor product operations composition, while (Clark et al., 2008) complements the previous approach with pregroup semantics. Similarly, Mitchell and Lapata (2008) employs vector sums and products, whereas Guevara (2010), Guevara (2011) and Baroni and Zamparelli (2010) model composition as a learnable function of two vectors. In the same vein, Paperno et al. (2014) proposes a generalised representation of composition functions.

At the same time, existing studies cover a wide range of modifier phenomena: adjective-noun (AN) compositions (Boleda et al., 2013), verb-argument composition (Lenci, 2011), determiner-noun (DP) phrases (Bernardi et al., 2013), recursive adjectival modifications (Vecchi et al., 2013), reverse adjectival composition for phrase generation (Dinu and Baroni, 2014), pointwise mutual information (PMI) analysis over AN compositions (Paperno and Baroni, 2016), morpheme representation (Marelli and Baroni, 2015) and metaphorical sense modeling (Lazaridou et al., 2013; Gutierrez et al., 2016).

More recently, syntax-aware composition of dependency tree nodes is comprehensively addressed by Weir et al. (2016), with empirical results tying previous approaches together. This work is complemented by Gamallo (2021) using contextual representations from transformer models. Finally, Purver et al. (2021) proposes a dynamic syntax framework for unambiguous composition of sentences through incremental semantic parsing, which was evaluated with non-contextual representations.

After the advent of contextual transformer-based representation, the interest has shifted into *testing* for specific compositional behaviours. The majority of existing works on *metamorphic* testing of language models focus on checking simple behavioural rules at scale (Belinkov and Bisk, 2018). This procedure is sometimes referred to as *behavioural* testing, as in (Ribeiro et al., 2020).

For example, Ma et al. (2020) investigate fairness-related behaviours by measuring the model robustness to changes in the gender of nouns or addition of population-specific adjectives. Similarly, Sun and Zhou (2018) focus on multi-language machine translation and compare direct translations to multi-hop ones. Likewise, Tu et al. (2021) test the robustness of question-answer systems to changes in the given text. Finally, Manino et al. (2022) define higher-order metamorphic relations that simultaneously mutate multiple base inputs. Thanks to this, they can test the systematicity and transitivity of language models.

Our work centers on behavioural testing for an embedding-denotation analogy, attempting to address concerns (in the fresh context of contextual transformer-based representations) such as the limitation stated in (Kartsaklis, 2014): that compositions may describe spurious relations which result from expressiveness limitations, rather than modelling theoretical compositional behaviour.

6 Conclusion

In this paper, we presented a methodology for measuring the presence and consistency of compositional behaviour in existing language models (LMs), comprising a set of tests for consistency of metamorphic relations associated to adjectival modifier phenomena in adjective-noun phrases, from a Montagovian formalism perspective. Our approach can provide important insight on LMs capabilities and limitations beyond semantic relatedness/similarity, helping to shape expectations on their use in applications with higher safety/criticality/fairness requirements. Although the tests are limited in scope, they can be applied to any language embedding.

Our empirical evaluation results indicate that current neural language models do not behave consistently according to expected behavior from the formalisms, with regard to the evaluated intersective and subsective properties. Such results imply that current language models, given limited context, may not be capable of capturing the evaluated semantic properties of language, or that linguistic theories from Montagovian tradition are not matching the expected capabilities of distributional models.

The proposed methodology is intended to be a stepping stone which can pave the way to a better understanding of LLMs latent spaces. Nevertheless, to improve our understanding of LLMs compositional capabilities, it is necessary to examine the relationship between the observed representation properties and specific NLP downstream task performance. Specifically, the alignment of compositional semantics between inputs and expected outputs, e.g., in RAG, summarization or Question Answering. Those should be explored in future work.

Future work also includes expanding the scope of the tests to other linguistic properties and an investigation on the effect of measured consistency on the relevant downstream tasks (e.g., NLI, RAG).

Limitations

Having been designed as a set of measurements for quasi-symbolic analogy, the presented approach is not intended to demonstrate or prove the properties of the distributional models but rather to verify compliance to particular behaviours of interest.

The formal linguistic properties evaluated in this study are not sufficient nor intended to evaluate general compositionality, but are a fundamental part of a larger set of compositional phenomena, which includes verbal, nominal and even non-linguistic (e.g., arithmetic) composition.

Furthermore, while the Montagovian perspective of compositionality is highly relevant from the symbolic and verification standpoints, other theoretical frameworks can present different constraints regarding word and phrase interpretations and are worthy of exploration.

Acknowledgements

This work was partially funded by the EPSRC grant EP/T026995/1 entitled "EnnCore: End-to-End Conceptual Guarding of Neural Architectures" under Security for all in an AI enabled society, by the Swiss National Science Foundation (SNSF) project NeuMath (200021_204617), by the CRUK National Biomarker Centre, and supported by the Manchester Experimental Cancer Medicine Centre and the NIHR Manchester Biomedical Research Centre.

References

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1183–1193.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Raffaella Bernardi, Georgiana Dinu, Marco Marelli, and Marco Baroni. 2013. A relatedness benchmark to test the role of determiners in compositional distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 53–57.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in bert. *Cognitive Computation*, 13(4):1008–1018.

Gemma Boleda, Marco Baroni, Louise McNally, et al. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013): long papers; 2013 Mar 20-22; Postdam, Germany. Stroudsburg (USA): Association for Computational Linguistics (ACL); 2013. p. 35-46.* ACL (Association for Computational Linguistics).

Danilo Silva Carvalho and Minh Le Nguyen. 2017. Building lexical vector representations from concept definitions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 905–915.

Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, T. H. Tse, and Zhi Quan Zhou. 2018. Metamorphic testing: A review of challenges and opportunities. *ACM Comput. Surv.*, 51(1).

Stephen Clark, Bob Coecke, and Mehrnoosh Sadrzadeh. 2008. A compositional distributional model of meaning. In *Proceedings of the Second Quantum Interaction Symposium (QI-2008)*, pages 133–140. Oxford.

Stephen Clark and Stephen Pulman. 2007. Combining symbolic and distributional models of meaning. In *Proceedings of AAAI Spring Symposium on Quantum Interaction*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*.

Georgiana Dinu and Marco Baroni. 2014. How to make words with vectors: Phrase generation in distributional semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 624–633.

R.M.W. Dixon, A.I. Aikhenvald, A.I.U. Aĭkhenval'd, and A.Y. Aikhenvald. 2004. *Adjective Classes: A Cross-Linguistic Typology*. Explorations in Language and S. OUP Oxford.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Deborah Ferreira, Julia Rozanova, Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2021. Does my representation capture x? probe-ably. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 194–201.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Pablo Gamallo. 2021. Compositional distributional semantics with syntactic dependencies and selectional preferences. *Applied Sciences*, 11(12):5743.

Emiliano Raul Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 workshop on geometrical models of natural language semantics*, pages 33–37.

Emiliano Raul Guevara. 2011. Computing semantic compositionality in distributional semantics. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

E Dario Gutierrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 183–193.

DunZhang [@HuggingFace]. 2024. Stella-en-1.5b-v5. https://huggingface.co/dunzhang/stella_en_1.5B_v5.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. 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)*, pages 4129–4142, Hong Kong, China. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Dimitri Kartsaklis. 2014. Compositional operators in distributional semantics. *Springer Science Reviews*, 2(1):161–177.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1517–1526.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Alessandro Lenci. 2011. Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd workshop on cognitive modeling and computational linguistics*, pages 58–66.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of distributional semantic models. *Language Resources and Evaluation*, pages 1–45.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Pingchuan Ma, Shuai Wang, and Jin Liu. 2020. Metamorphic testing and certified mitigation of fairness violations in NLP models. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 458–465. ijcai.org.

Edoardo Manino, Julia Rozanova, Danilo Carvalho, Andre Freitas, and Lucas Cordeiro. 2022. Systematicity, compositionality and transitivity of deep NLP models: a metamorphic testing perspective. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2355–2366, Dublin, Ireland. Association for Computational Linguistics.

Marco Marelli and Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological review*, 122(3):485.

Alessio Miaschi and Felice Dell'Orletta. 2020. Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *proceedings of ACL-08: HLT*, pages 236–244.

M. Morzycki. 2016. *Modification*. Key Topics in Semantics and Pragmatics. Cambridge University Press.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874.

OpenAI. 2024. Embeddings. https://platform.openai.com/docs/guides/embeddings/.

Denis Paperno and Marco Baroni. 2016. When the whole is less than the sum of its parts: How composition affects pmi values in distributional semantic vectors. *Computational Linguistics*, 42(2):345–350.

Denis Paperno, Marco Baroni, et al. 2014. A practical and linguistically-motivated approach to compositional distributional semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 90–99.

Ellie Pavlick and Chris Callison-Burch. 2016. So-called non-subsective adjectives. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 114–119, Berlin, Germany. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Mohammad Taher Pilehvar and Roberto Navigli. 2015. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228:95–128.

Matthew Purver, Mehrnoosh Sadrzadeh, Ruth Kempson, Gijs Wijnholds, and Julian Hough. 2021. Incremental composition in distributional semantics. *Journal of Logic, Language and Information*, 30(2):379–406.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*, pages 3982–3992.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Pia Sommerauer and Antske Fokkens. 2019. Conceptual change and distributional semantic models: An exploratory study on pitfalls and possibilities. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 223–233.

Liqun Sun and Zhi Quan Zhou. 2018. Metamorphic testing for machine translations: Mt4mt. In *2018 25th Australasian Software Engineering Conference (ASWEC)*, pages 96–100.

Kaiyi Tu, Mingyue Jiang, and Zuohua Ding. 2021. A metamorphic testing approach for assessing question answering systems. *Mathematics*, 9(7).

Eva Maria Vecchi, Roberto Zamparelli, and Marco Baroni. 2013. Studying the recursive behaviour of adjectival modification with compositional distributional semantics. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 141–151.

David Weir, Julie Weeds, Jeremy Reffin, and Thomas Kober. 2016. Aligning packed dependency trees: a theory of composition for distributional semantics. *Computational Linguistics*, 42(4):727–761.

A List of adjectives and nouns

Subsective (Intersective): wild, red, Canadian, depressed, square, seasonal, flamboyant, vigorous, loud, orange, shy.
Synonyms: feral, crimson, North American, melancholic, cuboid, periodic, exuberant, robust, cacophonous, peach, timid.

Subsective (Non-Intersective): skilful, powerful, particular, extreme, rare, unexpected.
Synonyms: skilled, potent, specific, severe, uncommon, surprising.

Plain Non-Subsective: former, alleged, apparent, arguable, assumed, believed, disputed, doubtful, erroneous, expected, faulty, future, historic, impossible, improbable, likely, ostensible, plausible, potential, proposed, putative, questionable, so-called, suspicious, theoretical, uncertain, unsuccessful.
Synonyms: previous, suspected, seeming, debatable, presumed, assumed, doubted, dubious, mistaken, predicted, broken, upcoming, legendary, unachievable, unlikely, probable, apparent, possible, possible, suggested, supposed, dubious, commonly-named, dubious, philosophical, tentative, failed
Privative Non-Subsective: artificial, counterfeit, deputy, ex-, fabricated, fictional, hypothetical, imaginary, mock, mythical, past, phony, spurious, virtual.
Synonyms: fake, forged, vice, former, forged, fictitious, supposed, imagined, simulated, fantastical, prior, fake, bogus, simulated.

Ambiguous: old, small, big.
Synonyms: aged, tiny, large.

Nouns: student, dog, potato, story, king, person, chair, occurence, law, problem, disaster, statement
Synonyms: learner, canine, tater, narrative, monarch, human, seat, happening, regulation, difficulty, catastrophe, declaration.

B Full experimental results

Tables 6 and 7 present the complete results of the set distance experiments for single phrases and phrase pairs, respectively (Section 4.2.1). Table 8 presents the complete results for the phrase-word (non-subsectivity) distance experiment (Section 4.2.3).

Models	Adjective Type Pair								
	(S-I, S-I)	(S-NI, S-I)	(NS-Pl, S-I)	(NS-Pr, S-I)	(A, S-I)	(S-I, S-NI)	(S-NI, S-NI)	(NS-Pl, S-NI)	(NS-Pr, S-NI)
DPR	0.5242	0.4268	0.5314	0.5222	0.6288	0.3939	0.3694	0.4347	0.3839
LaBSE	0.9205	0.9343	0.9534	0.9139	0.9671	0.9431	0.9555	0.9295	0.9394
Specter	0.6667	0.7273	0.7177	0.6661	0.7348	0.7955	0.8861	0.8184	0.7857
TE3-small	0.9773	0.9886	0.9941	0.9973	0.9899	0.9760	0.9583	0.9820	0.9901
NV-Embed-v2	0.7788	0.7134	0.6838	0.8139	0.7475	0.7437	0.5583	0.4892	0.7123
stella_en_1.5B_v5	0.9992	0.9949	0.9992	1.0	1.0	0.9962	1.0	0.9990	0.9980
Glove	1.0	1.0	1.0	0.9393	1.0	1.0	1.0	1.0	0.9414
Word2Vec	0.9969	1.0	0.9691	0.9404	1.0	1.0	1.0	0.9686	0.9394
	(A, S-NI)	(S-I, NS-Pl)	(S-NI, NS-Pl)	(NS-Pl, NS-Pl)	(NS-Pr, NS-Pl)	(A, NS-Pl)	(S-I, NS-Pr)	(S-NI, NS-Pr)	(NS-Pl, NS-Pr)
DPR	0.5324	0.4571	0.3529	0.3681	0.3997	0.5401	0.5536	0.4345	0.4782
LaBSE	0.9814	0.9584	0.9274	0.8360	0.8858	0.9588	0.8874	0.9176	0.8536
Specter	0.8194	0.7932	0.8009	0.6899	0.7235	0.7963	0.7002	0.7380	0.6960
TE3-small	0.9769	0.9823	0.9784	0.9160	0.9625	0.9609	0.9729	0.9692	0.9372
NV-Embed-v2	0.7222	0.6512	0.4789	0.3251	0.5223	0.5895	0.8236	0.7292	0.5686
stella_en_1.5B_v5	1.0	0.9992	0.9974	0.9771	0.9894	1.0	0.9984	0.9931	0.9850
Glove	1.0	1.0	1.0	0.9992	0.9442	1.0	0.9393	0.9414	0.9442
Word2Vec	1.0	0.9691	0.9404	0.9394	0.9122	0.9691	0.9404	0.9394	0.9122
	(NS-Pr, NS-Pr)	(A, NS-Pr)	(S-I, A)	(S-NI, A)	(NS-Pl, A)	NS-Pr, A)	(A, A)		
DPR	0.5069	0.6409	0.6263	0.5278	0.6029	0.5734	0.5139		
LaBSE	0.6483	0.9166	0.9848	0.9861	0.9701	0.9523	1.0		
Specter	0.6662	0.7619	0.7753	0.8519	0.7891	0.7361	0.6806		
TE3-small	0.9139	0.9583	0.9798	0.9491	0.9856	0.9722	0.9722		
NV-Embed-v2	0.6346	0.7679	0.7879	0.7639	0.7006	0.8115	0.4861		
stella_en_1.5B_v5	0.9547	0.9921	1.0	0.9954	1.0	1.0	1.0		
Glove	0.8873	0.9345	1.0	1.0	1.0	0.9345	1.0		
Word2Vec	0.8736	0.9404	1.0	1.0	0.9691	0.9404	1.0		

Table 6: Satisfaction score (consistency) for the set distance property (Equation 4), for noun phrases with pairs of adjectives of the indicated types (AAN format). We use the following shorthand notation in the table columns: A: Ambiguous, S-I: Subsective-Intersective, S-NI: Subsective Non-Intersective, NS-Pl: Plain Non-Subsective, NS-Pr: Privative Non-Subsective

Models	Adjective Type Pair								
	(S-I, S-I)	(S-I, S-NI)	(S-I, NS-Pl)	(S-I, NS-Pr)	(S-I, A)	(S-NI, S-NI)	(S-NI, NS-Pl)	(S-NI, NS-Pr)	(S-NI, A)
DPR	0.5000	0.3216	0.3380	0.4999	0.4238	0.5000	0.4882	0.6695	0.5834
LaBSE	0.5000	0.4252	0.3386	0.5268	0.3316	0.5000	0.4155	0.6216	0.3877
Specter	0.5000	0.6530	0.5461	0.5042	0.5735	0.5000	0.4091	0.3710	0.4014
TE3-small	0.5000	0.5108	0.4824	0.4822	0.8223	0.5000	0.4527	0.4552	0.8723
NV-Embed-v2	0.5000	0.5412	0.5148	0.5158	0.8230	0.5000	0.4769	0.4825	0.8759
stella_en_1.5B_v5	0.5000	0.7597	0.6379	0.5782	0.9090	0.5000	0.3897	0.3241	0.6614
Glove	0.5000	0.6565	0.6940	0.7022	0.4684	0.5000	0.5428	0.5552	0.3215
Word2Vec	0.5000	0.7536	0.6518	0.4879	0.9982	0.5000	0.4289	0.2953	0.8103
	(NS-Pl, NS-Pl)	(NS-Pl, NS-Pr)	(NS-Pl, A)	(NS-Pr, NS-Pr)	(NS-Pr, A)	(A, A)			
DPR	0.5000	0.6656	0.5971	0.5000	0.4487	0.5000			
LaBSE	0.5000	0.6838	0.5209	0.5000	0.2902	0.5000			
Specter	0.5000	0.4595	0.5145	0.5000	0.5318	0.5000			
TE3-small	0.5000	0.5001	0.8100	0.5000	0.8334	0.5000			
NV-Embed-v2	0.5000	0.5016	0.8686	0.5000	0.9053	0.5000			
stella_en_1.5B_v5	0.5000	0.4414	0.7617	0.5000	0.8252	0.5000			
Glove	0.5000	0.5087	0.2626	0.5000	0.2558	0.5000			
Word2Vec	0.5000	0.3817	0.8503	0.5000	0.8642	0.5000			

Table 7: Satisfaction score (consistency) for the set distance property across adjective phrase pairs (Equation 5), for noun phrases with pairs of adjectives of the indicated types (AN format). We use the following shorthand notation in the table columns: A: Ambiguous, S-I: Subsective-Intersective, S-NI: Subsective Non-Intersective, NS-Pl: Plain Non-Subsective, NS-Pr: Privative Non-Subsective

Models	Adjective Type				
	Subsective (Intersective)	Subsective (Non-Intersective)	Non-Subsective (Plain)	Non-Subsective (Privative)	Ambiguous
DPR	0.4621	0.3750	0.4784	0.5357	0.3889
LaBSE	0.3560	0.3055	0.5123	0.3273	0.1944
Specter	0.4848	0.3056	0.4907	0.5714	0.3333
TE3-small	0.8106	0.7500	0.7438	0.7738	0.3889
NV-Embed-v2	0.8409	0.7917	0.7901	0.8274	0.8056
stella_en_1.5B_v5	0.8106	0.5556	0.5772	0.6369	0.3333
Glove	0.6060	0.2222	0.2191	0.3214	0.2777
Word2Vec	0.5530	0.2083	0.3364	0.4940	0.0

Table 8: Non-subsectivity experiment, reporting satisfaction score (consistency) for the property in equation 7, as its expectation for the phrase dataset (Equation 8).