Reference: Benbaji-Elhadad, Ido, Omri Doron & Adele Hénot-Mortier. 2023. Distinguishing levels of morphological derivation in word-embedding models. In: Suet-Ying Lam and Satoru Ozaki (eds.), NELS 53: Proceedings of the Fifty-Third Annual Meeting of the North East Linguistic Society, Vol. 1, 63-76. Amherst, MA: GLSA.

Distinguishing levels of morphological derivation in word-embedding models*

Ido Benbaji-Elhadad, Omri Doron & Adèle Hénot-Mortier

Massachusetts Institute of Technology

1. Introduction

Word-embedding models, a family of computational models of the lexicon that assign lexical items to dense, high-dimensional vector representations of their meanings, have become immensely popular in the field of natural language processing (see, a.o., Mikolov et al. 2013a,b, Pennington et al. 2014, Zhao et al. 2019). These models instantiate the *distributional hypothesis* in linguistics, according to which a systematic co-occurrence of two words in similar contexts is indicative of a similarity in the meanings of these two words (Harris 1954, Firth 1957). This is due to the fact that word-embeddings represent the meaning of a word X as a vector whose values are determined, roughly, by the other words that co-occur with X in a corpus (Jurafsky and Martin 2000).

The current study provides evidence that these models are sensitive to a distinction in the study of word-formation between two levels of morphological derivation. Since at least Siegel 1974, morphological processes have been organized in a hierarchy of two levels, each associated with a distinctive phonology and semantics: a lower level (henceforth, LL), where derivations may be irregular, non-productive and semantically unpredictable, and an upper one (henceforth, UL) where they are regular and generally more productive and predictable. Effectively, the two-level hypothesis posits that words derived by certain morphological processes, diagnosed via their phonological effect, show certain semantic systematicities that words derived via other processes do not. If the vectors in word-embedding models represent *meanings* of words, as they purportedly do, we expect these models to capture the semantic differences observed in the morphological literature *vis a vis* the two-level hypothesis. This paper provides a proof of concept that, indeed, they do.

For a concrete example of the two-level architecture of morphological derivation, take the English suffixes /*-ity*/ and /*-ness*/. The two both join an adjectival base to form an abstract noun. However, the phonological effect brought about by each suffix is different:

[*] Thanks to Adam Albright, Michael Elhadad, Itamar Kastner and Roger Levy for fruitful comments. The code for this project is available on GitHub at the following address: `https://github.com/AdeleMortier/appendix_NELS53_BE_D_HM`.

only /-*ity*/ shifts the stress of the input to suffixation to the syllable preceding the suffix, while /-*ness*/ leaves the stress-assignment of its input intact, (Chomsky and Halle 1968, Siegel 1974, Aronoff 1976, Kiparsky 1982). As the examples in (1) illustrate, while /-*ity*/ seems to have access to the stress features of its base, /-*ness*/ seems to lack similar access.

(1)	áctive	actívity	áctiveness
	mónstrous	monstrósity	mónstrousness
	équal	eqúality	équalness
	fátal	fatálity	f́atalness

Aronoff (1976) observes that this phonological difference coincides with a semantic one: while the meaning that is derived by suffixation with /-*ness*/ is fully predictable given the meaning of its input, this is not the case for /-*ity*/. Attaching /-*ness*/ to an adjectival base *X* results in nouns denoting the quality of being *X*, the extent to which something is *X*, or the fact that something is *X*. The nouns that results from /-*ity*/ suffixation are related to their adjectival input in some sense, but the relation is not predictable. For instance, *monstrosity* can denote a large and unsightly building, and *fatality* can denote death in an accident, at least in addition to the meanings that result from suffixation with with /-*ness*/.

According to the two-level hypothesis, the difference between /-*ity*/ and /-*ness*/ stems from the fact that suffixation with each of them involves a different level of morphological derivation: attaching /-*ity*/ to a base is a LL derivation, while attaching /-*ness*/ is an UL one. More generally, given two words derived from a base X, the hypothesis argues that the phonological and semantic relation between them depends on the level in which their derivations diverge; given derivatives X' and X'', where $X \rightarrow X'$ is a LL derivation (e.g., *monstrous−ity*) and $X \rightarrow X''$ is an UL one (e.g., *monstrous−ness*), the hypothesis predicts a more arbitrary semantic relation between X' and X than between X'' and X, correlating with a difference in the derivations' phonological effect on X.

This paper uses the case of English suffixation to illustrate that vector representations of derived words and their bases in word-embedding models do, indeed, conform to the tenets of the two-level hypothesis. To do so, in section 2, we first provide a general introduction to word-embedding models, and introduce quantitative methods for evaluating semantic predictability within such models. We use these methods in section 3, which illustrates that the vectors that different word-embedding models assign to instantiations of LL English suffixes in a corpus are less semantically predictable than those assigned to instantiations of UL suffixes. Section 4 reviews our previous attempt to demonstrate the sensitivity of word-embeddings to the two-level distinction in Benbaji et al. (2022), and points to some crucial shortcomings that render that study insufficient. Section 5 concludes.

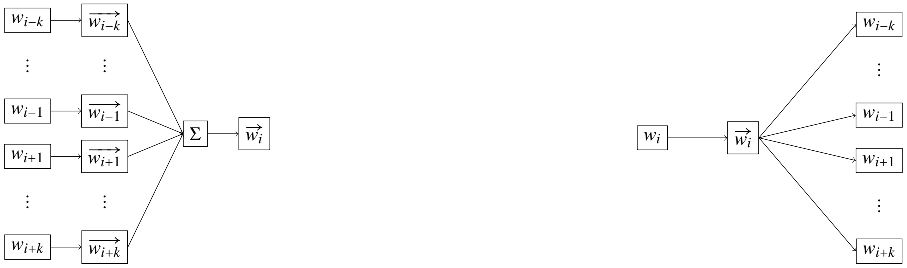
2. Word-embeddings and quantitative measures of semantic predictability

A word-embedding is a representation of a word in the form of a dense, high-dimensional, real-valued vector. It is designed to encode the meaning of a word in such a way that words that are closer in the vector space are expected to be similar in meaning (Jurafsky and Martin 2000). To achieve this goal, word-embeddings often end up associating each dimension of the vector space with a tangible semantic or distributional feature (e.g. ±human, ±feminine, ±function word, ±verb) – or potentially, a combination thereof.

How are efficient and meaningful word-embeddings derived? Popular methods include neural networks (Mikolov et al. 2013b) and dimensionality reduction performed on a word co-occurrence matrix (Lebret and Collobert 2013). The models examined in this paper – GloVe, Word2Vec, fastText, BERT – employ different methods, but the general ingredients remain the same: a model is passed an unstructured input and learns a condensed internal representation of this input while optimizing a language-related objective function.

In the case of GloVe (Pennington et al. 2014), the vector representation is obtained from the (preprocessed) word co-ocurrence matrix, with the objective that the dot product of any two word-vectors (a measure of their similarity) be equal to the logarithm of the corresponding words' probability of co-occurrence. In the Word2Vec (Mikolov et al. 2013a,b) and fastText (Bojanowski et al. 2016) models, the representation is learned from one-hot encoded vectors[1] or combinations thereof, by either predicting any target word from the representation of the words surrounding it (Continuous Bag Of Words or CBOW cf. Figure (2a)), or, by predicting the context from the target word (Skip-gram, cf. Figure (2b)).[2]

(2) *Basic architectures used by fastText and Word2Vec to derive word-embeddings*



a. The CBOW architecture

b. The Skip-gram architecture

Unlike Word2Vec, fastText models a word as the mean of its constituting *n*-grams – making the model more sensitive to morphological regularities. This design also ensures that a vector can be obtained from any given word, even one that has not been previously "seen" by the model, which is why embeddings like fastText are sometimes called *dynamic*, by contrasts with *static* models such as Word2Vec and GloVe. BERT finally, adopts the more recent Transformer architecture (Vaswani et al. 2017). It is trained to perform masked language modeling (predict a masked word given a context of arbitrary size), and next-sentence prediction. BERT embeddings are also dynamic, and were obtained in our case by extracting the weights of the second-to-last layer of the model (following Xiao 2018).

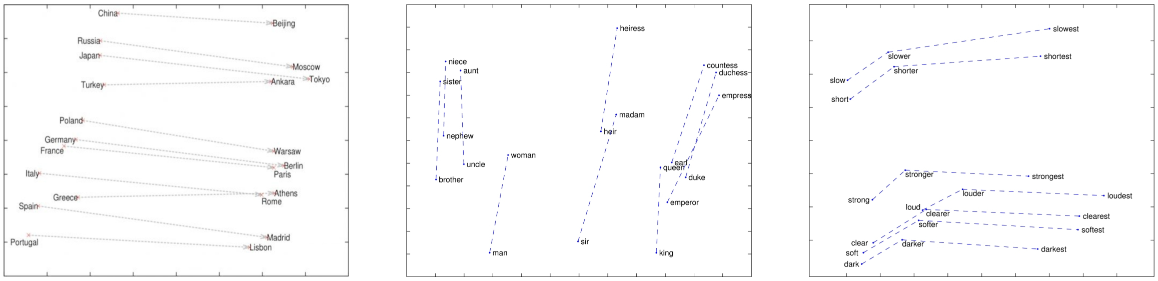
As mentioned above, word-embeddings are derived as byproducts of a given learning task; the assumption being that to perform well at such tasks, given an embedding space

[1] The one-hot vector for the *i*-th element of a lexicon is defined as the vector whose components are all zero except the *i*-th component, which is equal to one.

[2] The Word2Vec models trained for this paper were initialized with the Skip-gram objective; the fastText models, with the CBOW objective.

whose size is significantly smaller than that of the input, a model needs to efficiently encode a certain number of distributional regularities of the language, pertaining to syntax, semantics, and, under certain conditions, morphology. While there is no formal guarantee that word-embeddings will capture correct linguistic generalizations (though see discussion in Allen and Hospedales 2019, Ethayarajh et al. 2019), empirical observations suggest that such models do, in fact, encode semantic relations such as *be the capital city of* (Mikolov et al. 2013b), features like gender, and aspects of morphological derivation like comparative and superlative formation (Pennington et al. 2014) – as illustrated in Figure (3).

(3) *2D reductions of the original GloVe/Word2Vec embeddings showing how semantic/morphological relationships are encoded into stable vector translations*



a. Country/Capital city b. Masculine/Feminine c. Comparative/Superlative

2.1 Stability of vector sets as a measure of semantic predictability

Consider Figure (3c), representing the position of vectors assigned to adjectives in their positive form ($\overrightarrow{\text{A}}$), relative to that of vectors assigned to their respective comparative ($\overrightarrow{\text{A-er}}$) and superlative ($\overrightarrow{\text{A-est}}$) forms. Each word in the figure represents a vector in the embedding space. The dotted lines linking positive-forms to their comparatives, and comparative-forms to their superlatives also represent vectors, i.e., those derived by subtracting positive-form vectors from comparative ones ($\{\overrightarrow{\text{A-er}} - \overrightarrow{\text{A}} \mid \text{A}\}$), and comparative-form vectors from superlative ones $\overrightarrow{\text{A-est}}$ ($\{\overrightarrow{\text{A-est}} - \overrightarrow{\text{A-er}} \mid \text{A}\}$). While these vectors can also be represented as points in the embedding space, they are depicted as lines here to illustrate their role as the "positive to comparative" and "comparative to superlative" transformations, respectively. These transformations seem remarkably "stable" – in the reduced space they are almost parallel! But how can stability be quantitatively defined over (multidimensional) vectors? We introduce next two measures of vector-set stability that will make use of a typical measure of similarity in a high-dimensional space; i.e., *cosine similarity*.

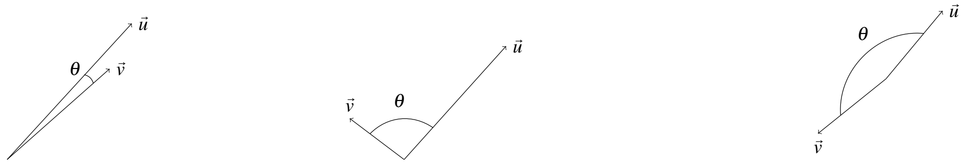
Cosine similarity is the cosine of the angle between two vectors being compared – itself equal to the dot product of the vectors divided by the product of their lengths (Equation A).

$$\mathcal{S}(\vec{u},\vec{v}) = \frac{\vec{u}.\vec{v}}{\|\vec{u}\| \times \|\vec{v}\|} \in [-1;+1] \tag{A}$$

It is a symmetric measure, as its value does not depend on the order of its arguments. As Figure (4) shows, it is also insensitive to differences of magnitude between the vectors;

the respective *proportions* of the different features (or vector dimensions) are considered when comparing two word-vectors, not the absolute values of those features.

(4) *Maximum cosine similarity is achieved when the vectors align (have the same direction), while minimum similarity corresponds to vectors with opposite directions*



a. θ ~ 0° $\mathcal{S}(\vec{u},\vec{v}) \sim +1$

b. θ ~ 90° $\mathcal{S}(\vec{u},\vec{v}) \sim 0$

c. θ ~ 180° $\mathcal{S}(\vec{u},\vec{v}) \sim -1$

With this in mind, we can define two measures of vector-set stability that we call *dispersion* and *variation*. Dispersion is the set of pairwise negative cosine similarities between all vectors in a set (cf. Equation B), which quantifies how the directions of all these vectors differ from each other. For a set with n vectors, there are $\frac{n(n-1)}{2}$ dispersion measures. Variation refers to the set of negative cosine similarities between each vector of a set and the centroid (mean vector) of the set (cf. Equation C). It quantifies how the directions of all vectors differ from that of their mean. A set with n vectors, has n variation measures.

$$Dispersion(\{\vec{v_1},\vec{v_2},\ldots,\vec{v_n}\}) = \{-\mathcal{S}(\vec{v_i},\vec{v_j}) \mid i > j\} \tag{B}$$

$$Variation(\{\vec{v_1},\vec{v_2},\ldots,\vec{v_n}\}) = \{-\mathcal{S}(\vec{v_i},\vec{v}) \mid \vec{v_i}\} \text{ with } \vec{v} \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n}\vec{v_i} \tag{C}$$

(5) *Graphical illustration of the measures of dispersion and variation (angles represent measures of cosine similarity)*

a. Dispersion

b. Variation (yellow vector represents the mean)

Dispersion and *Variation* both map a set of multidimensional vectors to a set of scalars, whose mean or median can inform us about the overall stability of the vector sample; the lower the dispersion and variation medians of a vector set are, the stabler that set is. For concreteness, consider Figure (3c) again, where the comparative transformation vectors seem almost parallel and have the same direction. The angle between any two of these vectors is thus approximately 0°, and the resulting cosine similarity is close to 1 (cf. Figure (4)). This in turn entails that the corresponding samples of dispersion and variation measures is associated to a mean/median close to −1 (cf. Equations B-C). The same holds for the superlative set. These are thus stable transformations.

Crucially, this notion of stability corresponds to what we typically think of as semantic predictability. A transformation is semantically predictable, if the meaning of its output is determined solely by the meaning of its input combined with the fixed semantic contribution of the transformation. In an embedding model, a stable transformation corresponds to a set of aligned vectors, whose shared direction corresponds to the fixed semantic contribution of the transformation. If this direction is known, then given an input word-vector, we can predict the meaning of the output word-vector without any further information.

Note, finally, that comparing the degree of stability between two vectors samples can be done through statistical hypothesis testing between the two corresponding samples of variation and dispersion measures (as will become relevant below). Both measures however, have some limitations. Variation is relatively robust with respect to outliers,[3] but may become unstable if the vectors are uncorrelated, as in that case the mean vector will become close to the null vector (for which cosine similarity cannot be computed). Dispersion is slightly more sensitive to outliers[4] and may artificially inflate the sample size when performing statistical testing, but has the advantage of yielding robust, well-defined measures.

3. Case study: English suffixation

As discussed in section 1, English suffixes come in two kinds: those that trigger stress-shifting, whose semantic contribution is not predictable from the meaning of their input, and those that leave stress-assignment untouched, whose semantics is fully predictable given that of their input. */-ity/* belongs to the former kind, while */-ness/* belongs to the latter, as shown in (1). Another two suffixes that exemplify the distinction are */-al/* and */-less/*: the former shifts the stress of its input while the latter does not – see (6).

(6)	cómmerce	commércial	cómmerceless
	préjudice	prejudícial	préjudiceless
	rémedy	remédial	rémediless
	séntiment	sentiméntal	séntimentless

It is easy to see that in this case, too, stress-shifting correlates with semantic unpredictability. While the X-*less* forms always denote the property of *lacking X*, */-al/* brings about a range of different meanings: something *commercial* is intended to make a profit, something *prejudicial* has a harmful, unfair influence, something *remedial* is provided as a remedy, and someone *sentimental* holds specific sentiments, i.e., tenderness and nostalgia.

We use the notion of stability in a vector-space introduced in section 2, to test whether word-embedding models indeed capture the difference in semantic effect of the two kind of suffixes. Lower level suffixes, like */-ity/* and */-al/* whose semantic effect is unpredictable, should correspond to vectors that are less stable (*higher* dispersion and variation) than the vectors corresponding to semantically predictable suffixes like */-ness/* and */-less/*.

[3] If a set of n vectors contains one outlier, only $nicefrac1n$ measures of variation will be influenced by it.

[4] If a set of n vectors contains one outlier, then $\frac{2(n-1)}{n(n-1)} = {}^{2}\!/_{n}$ measures of variation will be influenced by it.

To test this hypothesis, we scanned an English dictionary (240788 tokens) for word triplets $\langle b, b\text{-}s_{LL}, b\text{-}s_{UL}\rangle$, where b is a base (resp. adjectival or nominal), $b\text{-}s_{LL}$ is the base adjoined to a LL suffix (resp. /*-ity*/ and /*-al*/), and $b\text{-}s_{UL}$ is the base followed by an UL suffix (resp. /*-ness*/ and /*-less*/). The resulting triplets were manually inspected to ensure well-formedness. We obtained the corresponding word-vector triplets $\left\langle \overrightarrow{b}, \overrightarrow{b\text{-}s_{LL}}, \overrightarrow{b\text{-}s_{UL}}\right\rangle$, in four pre-trained embedding models – BERT (Devlin et al. 2018), FastText (Bojanowski et al. 2016, Grave et al. 2018), GloVe (Pennington et al. 2014) and Word2Vec (Mikolov et al. 2013a,b) – and reduced their dimensions using Principal Component Analysis (PCA), cf. Table 1.[5] We employ a method recently used by Marelli and Baroni (2015), Hénot-Mortier (2022), Naranjo and Bonami (2023), Bonami and Naranjo (2023) to compute, for each triplet, the corresponding "affixal" vectors, derived by subtracting the vector of the base from that of the affixed form, as in Equations D-E.

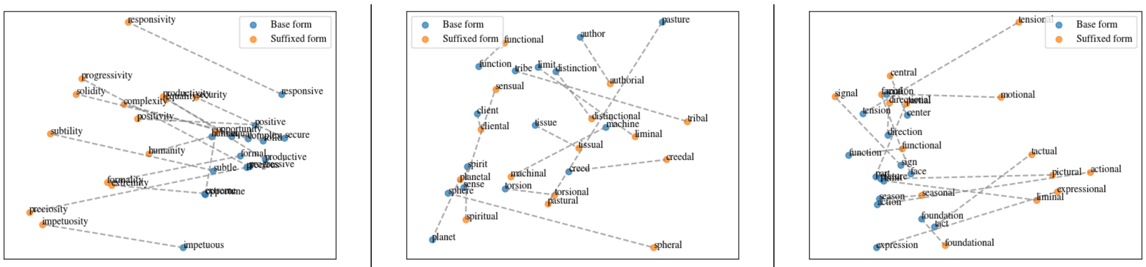
Model	Initial Dimension	Reduced dimension	
		(*-ity*, *-ness*)	(*-al*, *-less*)
BERT	768	198	152
FastText	300	130	84
GloVe	300	129	79
Word2vec	300	52	32

Table 1: Dimensions before/after reduction.

$$\forall \left\langle \overrightarrow{b}, \overrightarrow{b\text{-}s_{LL}}, \overrightarrow{b\text{-}s_{UL}}\right\rangle : \qquad \overrightarrow{s^{b}_{LL}} = \overrightarrow{b\text{-}s_{LL}} - \overrightarrow{b} \tag{D}$$

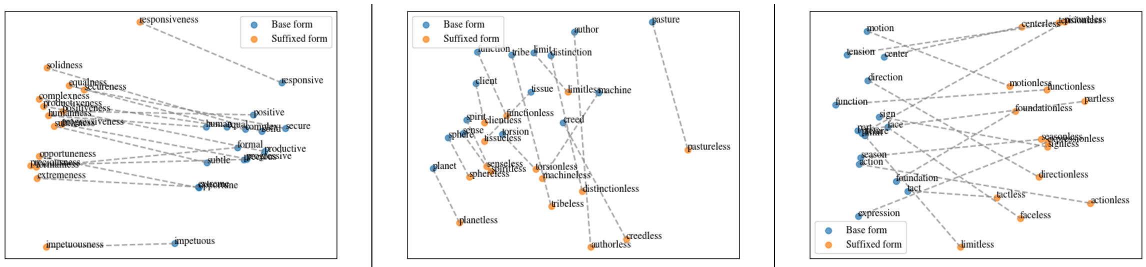
$$\overrightarrow{s^{b}_{UL}} = \overrightarrow{b\text{-}s_{UL}} - \overrightarrow{b} \tag{E}$$

(7) *2D plots (cosine kernel PCA) of samples of 15 pairs of word vectors (base forms and corresponding affixed form). "Affixal" vectors appear as dashed grey lines*



A→ A-*ity* (LL) N→ N-*al* (LL) N→ N-*al* (LL)

[5] PCA removed the dimensions irrelevant to our problem, while retaining 90% of the explained variance.



A→ A-*ness* (UL)

a. FastText

N→ N-*less* (UL)

b. BERT

N→ N-*less* (UL)

c. GloVe

Figure (7) gives an impression of the differences in vector stability between LL and UL vectors: the dotted lines, corresponding to affixal vectors, appear more aligned on the second row of figures (UL suffixes) than on the first (LL suffixes). Less impressionistically, of course, we can measure and compare vector-set stability using *dispersion* and *variation* introduced in section 2.

The hypothesis, recall, is that a sample of LL affixal vectors should exhibit less stability, i.e., more dispersion and more variation, than a comparable population of UL affixal vectors, in line with the two-level architecture of morphological derivation:

$$\forall (s_{LL}, s_{UL}) \in \{(\textit{/-ity/}, \textit{/-ness/}), (\textit{/-al/}, \textit{/-less/})\}:$$

$$Dispersion(\{\overrightarrow{s^b_{UL}} \mid b\}) \prec Dispersion(\{\overrightarrow{s^b_{LL}} \mid b\}) \quad \text{(F)}$$

$$Variation(\{\overrightarrow{s^b_{UL}} \mid b\}) \prec Variation(\{\overrightarrow{s^b_{LL}} \mid b\}) \quad \text{(G)}$$

In the above formulae, $\prec$ denotes a statistical inequality between two sets of measures, here tested using a two-sided (paired) Wilcoxon test. Effect sizes were computed in the form of Cliff's Δ (robust, non-parametric) and Cohen's *d* (standard, parametric). The hypothesis that UL transformations are more stable than LL ones is very consistently verified for the (*/-ity/*, */-ness/*) pair (cf. Table 2), and associated with medium to large effect sizes. The effect is not as large for the (*/-al/*, */-less/*) pair (cf. Table 3), especially regarding the variation metric. We think two factors might explain this mixed pattern. First, the lower sample sizes in that particular setting (only 15 datapoints could be tested with Word2Vec!) may have prevented small effects from reaching significance. Second, */-al/* and */-less/* differ more in their semantics than */-ity/* and */-ness/*, which may be a potential confound in the comparison of their variability.

Model	Dispersion							Variation						
	Sample Size	*p*-value		Cliff's Δ		Cohen's d		Sample Size	*p*-value		Cliff's Δ		Cohen's d	
BERT	185745	0	****	.30	(S)	.54	(M)	610	2.00e-59	****	.49	(L)	.89	(L)
FastText	10296	0	****	.38	(M)	.68	(L)	144	3.95e-13	****	.40	(M)	.72	(L)
GloVe	7875	0	****	.46	(M)	.82	(L)	126	4.73e-12	****	.56	(M)	.87	(L)
Word2vec	406	1.16e-10	****	.21	(S)	.38	(M)	29	7.60e-2	.	.27	(S)	.42	(M)

Table 2: Stability of $\overrightarrow{\textit{/-ity/}}$ vs. $\overrightarrow{\textit{/-ness/}}$. N, S, M, L resp. mean negligible, small, medium and large effect size. Green rows indicate significant results ($p < .05$).[6]

Model	Dispersion				Variation			
	Sample Size	*p*-value	Cliff's Δ	Cohen's d	Sample Size	*p*-value	Cliff's Δ	Cohen's d
BERT	20910	0 ****	.47 (M)	.88 (L)	205	9.05e-26 ****	.65 (L)	1.30 (L)
FastText	1431	1.22e-11 ****	.11 (N)	.24 (S)	54	1.70e-1	.16 (S)	.29 (S)
GloVe	1176	6.98e-108 ****	.53 (L)	.97 (L)	49	1.31e-9 ****	.65 (L)	1.32 (L)
Word2vec	105	5.44e-2 .	.12 (N)	.27 (S)	15	4.89e-1	.29 (S)	.41 (M)

Table 3: Stability of $\overrightarrow{\textit{/-al/}}$ vs. $\overrightarrow{\textit{/-less/}}$, same conventions as above.

3.1 Productive /-*ity*/ suffixes: initial results

Aronoff and Lindsay (2014) observe that the LL suffix /-*ity*/ demonstrates "upper level" behavior in two specific contexts: when it does not influence stress assignment, and when it evidently stacks over other suffixes such as /-*able*/. We focus here on the later case. To determine if /-*ity*/ in the context of /-*able*/ is indeed more regular semantically, we compared the population of the corresponding $\overrightarrow{\textit{/-ity/}}$ vectors (that we call here -*(abil)/*$\overrightarrow{\textit{ity/}}$ to avoid confusion) to that of of the previously studied affixal vectors ($\overrightarrow{\textit{/-ness/}}$ and $\overrightarrow{\textit{/-ity/}}$). We expect the -*(abil)/*$\overrightarrow{\textit{ity/}}$ vectors to be as stable as the $\overrightarrow{\textit{/-ness/}}$ vectors, and overall more stable than the other $\overrightarrow{\textit{/-ity/}}$ vectors.[7] Table 4 shows that the stability measure of -*(abil)/*$\overrightarrow{\textit{ity/}}$ is intermediate between that of $\overrightarrow{\textit{/-ity/}}$ and that of $\overrightarrow{\textit{/-ness/}}$. Note, however, that for BERT and Word2Vec, the effect size associated with the -*(abil)/ity*/ vs. /-*ness*/ comparison is fairly small, and smaller than the effect size of the -*(abil)/ity*/ vs. /-*ity*/ comparison – weakly suggesting that the semantic predictability of -*(abil)/ity*/ may be closer to that of /-*ness*/ than to that of /-*ity*/. Those results however come with a caveat: contrary to the -*(abil)/*$\overrightarrow{\textit{ity/}}$ sample, the $\overrightarrow{\textit{/-ness/}}$ and $\overrightarrow{\textit{/-ity/}}$ samples used in the above comparison were linked to similar bases. Such a pairing is not possible for the -*(abil)/*$\overrightarrow{\textit{ity/}}$ sample, as English does not stack -*ness* over -*able*. This entails that the difference measured between the -*(abil)/*$\overrightarrow{\textit{ity/}}$ sample and the other sample might be due to some idiosyncratic behavior of words ending in -*able*, independent of the two-level model of morphology. We leave further discussion of this issue for future work.

[6] Significance levels were set as follows:: $p < 1\text{e-}4 \Rightarrow$ '****'; $p < 1\text{e-}3 \Rightarrow$ '***'; $p < 1\text{e-}2 \Rightarrow$ '**'; $p < 5\text{e-}2 \Rightarrow$ '*'; $p < 1\text{e-}1 \Rightarrow$ '.' . Effect size thresholds (in absolute value) for Cliff's Δ were: $[0;.147[\Leftrightarrow$ **N**egligible; $[.147;.33[\Leftrightarrow$ **S**mall; $[.33;.474[\Leftrightarrow$ **M**edium; $[.474;+\infty[\Leftrightarrow$ **L**arge. For Cohen's *d*: $[0;.1[\Leftrightarrow$ N; $[.1;.35[\Leftrightarrow$ S; $[.35;.65[\Leftrightarrow$ M; $[.65;+\infty[\Leftrightarrow$ L.

[7] The reason we avoid exploring non-stress-shifting cases of /-*ity*/ in embedding models is that this would involve finding bases in which stress is not final, which is not straightforward and may not yield enough data. Whenever the base carries final stress, one cannot tell if the derived /-*ity*/ form with antepenultimate stress is vacuously "shifted" – or not. Unambiguous bases thus usually contain 3 or more syllables.

Model	Comparison	Measure & Test[8]	
		Dispersion	Variation
BERT	-(*abil*)/*ity*/ vs. /-*ity*/	$\overset{****}{<}$ (S)	$\overset{****}{<}$ (M/L)
	-*ness* vs. -(*abil*)/*ity*/	$\overset{****}{<}$ (N/S)	$\overset{***}{<}$ (N/S)
fastText	-(*abil*)/*ity*/ vs. /-*ity*/	$\overset{****}{<}$ (S/M)	$\overset{****}{<}$ (S/M)
	-*ness* vs. -(*abil*)/*ity*/	$\overset{****}{<}$ (S/M)	$\overset{****}{<}$ (S/M)
GloVe	-(*abil*)/*ity*/ vs. /-*ity*/	$\overset{****}{<}$ (S/M)	$\overset{***}{<}$ (S/M)
	-*ness* vs. -(*abil*)/*ity*/	$\overset{****}{<}$ (L)	$\overset{****}{<}$ (L)
Word2Vec	-(*abil*)/*ity*/ vs. /-*ity*/	$\overset{****}{<}$ (M/L)	$\overset{***}{<}$ (M/L)
	-*ness* vs. -(*abil*)/*ity*/	$\overset{****}{<}$ (S)	$\dot{<}$ (S)

Table 4: Two-way comparisons between the stability of *-(abil)*$\overrightarrow{\textit{/ity/}}$ and resp. $\overrightarrow{\textit{/-ity/}}$ and $\overrightarrow{\textit{/-ness/}}$

4. Hebrew denominal verbs

In Benbaji et al. 2022, we attempted to examine the sensitivity of word-embedding models to the two-level architecture using a data set from Modern Hebrew. This section briefly reviews the results of that study, illustrating why, while suggestive, they do not justify the claim we argue for in this paper; namely, that word-embedding models are sensitive to the two-level distinction in morphology.

4.1 Background

In Benbaji et al. 2022 we used a dataset which involves Hebrew denominal verbs, namely verbs derived from nouns. We took for granted the well-known analysis of such verbs in Arad 2003, according to which they are the result of applying a nominalizing head to an abstract *root* (assumed to be an LL operation), and then applying a verbalizing head to the result (a UL operation). Our data set consisted of triplets of the form <N, V, D> such that N is a noun, V a verb or a group of verbs derived from the same root as N, and D a denominal verb derived from N. N and V are both assumed to derive from LL operations over abstract roots, while, as stated above, N and D are related to each other via an UL process.

What does the two-level hypothesis predict in this case? If one accepts Arad's analysis, according to which non-denominal verbs are derived via LL operations over abstract roots, while denominal ones involve UL operations over words, the hypothesis predicts N and D to be related by a stable transformation, while the relation between N and any root derived verb V should be arbitrary, as they are all derived via LL operations from abstract roots.

Evaluating the prediction in word-embeddings is not obvious. While it is certainly possible to measure the predictability of the UL step from N to D (using the stability measure developed above), it is difficult to compare that measure to the stability of the derivation of

[8] Mann–Whitney U test, Holm–Bonferroni correction for multiple hypothesis testing.

N and V from an abstract root, which lacks an independent representation in a corpus and thus lacks a straightforward vectoral representation.

In lieu of comparing stability measures, we interpreted the prediction of the two-level model as implying that D should be *semantically close* to N. More precisely, given an abstract root $\sqrt{}$, we define $Area(\sqrt{})$ as the convex envelope of the set of all words derived from that root $\{\overrightarrow{X} | \sqrt{} \longrightarrow^{*} X\}$.[9] We can formulate the prediction as follows:

(8) a. Given a root $\sqrt{}$, and words A, B, s.t. $\sqrt{} \xrightarrow{LL} A$, and $\sqrt{} \xrightarrow{LL} B$, we expect $\overrightarrow{A}$ and $\overrightarrow{B}$ to be randomly distributed across $Area(\sqrt{})$.

b. Given $\sqrt{}$, A and B, s.t. $\sqrt{} \xrightarrow{LL} A \xrightarrow{UL} B$, we expect $\overrightarrow{A}$ and $\overrightarrow{B}$ to be relatively close to each other within $Area(\sqrt{})$.

To measure proximity, we use the standard metric of cosine similarity (see equation A in section 2.1 above for a formal definition). It follows from (8) that for our triplets of the form <N, V, D>, D should be closer to N than V is. This is the hypothesis we set out to test.

Hebrew is a good testing ground for that hypothesis because, if Arad (2003) is correct, it provides us with an orthographically-represented diagnostic for denominal verbs (unlike, say, English): In short, Hebrew words are generally derived by applying a template (which determines the syntactic category of the word) to a three-consonant root (which never appears in the language without a template). Templates are discontinuous sequences of phonemes, consisting mostly (but not only) of vowels, with gaps that are filled by the root's consonants. Arad notices that some verbs contain a consonant belonging neither to their root nor to their verbal template, but to a nominal template. She argues that these verbs must involve denominalization; i.e., result from first applying the nominal template to a root, and only then applying the verbal template. We used this diagnostic to generate a pool of denominal verbs (along with the nouns they are derived from and a set of verbs derived from the same root).

4.2 Results

Triplets of the form $\left\langle N_{\sqrt{}}, V_{N_{\sqrt{}}}, \{V_{\sqrt{}}^{(1)}, \ldots, V_{\sqrt{}}^{(k)}\}\right\rangle$, where $\sqrt{}$ is a fixed root, $N_{\sqrt{}}$ a root-derived noun, $V_{N_{\sqrt{}}}$ a denominal verb derived from it, and $\{V_{\sqrt{}}^{(1)}, \ldots, V_{\sqrt{}}^{(k)}\}$ a set of root-derived verbs, were semi-automatically generated using the (morphologically annotated) Knesset Meetings Corpus (Itai and Wintner 2019), by matching words against predefined templates involving templatic consonants. Triplets were manually inspected to ensure they were well-formed and minimally ambiguous, resulting in 66 datapoints. Word vectors were obtained using four different word-embedding models: AlephBERT (Devlin et al. 2018, Seker et al. 2021), FastText (Bojanowski et al. 2016, Grave et al. 2018), GloVe (Pennington et al. 2014) and Word2Vec (Mikolov et al. 2013a,b); and various initial embedding dimensions. Aleph-

[9] $*$ denotes the reflexive transitive closure of a relation (here, that of morphological derivation).

BERT and FastText300 were pretrained; the other models were trained by us. Dimensionality was reduced using Principal Component Analysis (PCA), as in section 3.

For each triplet, the following inequality between cosine similarities was expected to hold in the reduced vector space:

$$\forall\sqrt{}:\ \frac{1}{k}\sum_{i=1}^{k}\mathcal{S}\left(\overrightarrow{N_{\sqrt{}}},\overrightarrow{V_{\sqrt{}}^{(k)}}\right) < \mathcal{S}\left(\overrightarrow{N_{\sqrt{}}},\overrightarrow{V_{N_{\sqrt{}}}}\right) \tag{H}$$

We used the same test as in section 3. The results in Table 5 show that the hypothesis is verified for most models and with overall large effect sizes. The BERT model seems to be the only true exception to an otherwise very robust pattern.

Model	Initial	Hypothesis H (mean)		
	Dim.	*p*-value	Cliff's Δ	Cohen's d
BERT	768	6.35e-2 .	-.23 (S)	-.37 (S)
FastText	50	4.99e-7 ****	.70 (L)	1.46 (L)
	100	2.79e-9 ****	.82 (L)	1.75 (L)
	300	4.66e-10 ****	.87 (L)	2.04 (L)
GloVe	50	1.07e-4 ***	.54 (L)	1.04 (L)
	100	1.54e-5 ****	.70 (L)	1.12 (L)
	300	1.95e-4 ***	.60 (L)	.96 (L)
Word2vec	50	4.00e-8 ****	.85 (L)	1.92 (L)
	100	4.00e-8 ****	.84 (L)	1.95 (L)
	300	6.52e-8 ****	.81 (L)	1.78 (L)

Table 5: Results of the tests, same conventions as in section 3.

4.3 Limitations of the Hebrew case study

This study has two crucial limitations that prevent it from constituting definitive evidence for the sensitivity of word-embeddings to the two-level distinction. First, our dataset construction relied on Arad's analysis of the templatic consonants on verbs as markers of denominalization. This however, has been challenged recently by Rasin et al. (2021), who argue that templatic consonants do not necessarily come from nominal bases, based on verbs in the *tiCCeC* template (e.g., **tiʃʔel**, 'interrogate'), which contain a templatic /*t*/ but lack a nominal counterpart.

A deeper issue lies in our translation of the two-level hypothesis' predictions into word-embedding models. Roughly, we assumed that the unpredictability of LL operations vs. the predictability of UL ones should translate into a contrast in *semantic similarity*,[10] but this assumption does not seem to us to be strongly rooted. A better way to test the two-level hypothesis would be to study differences in stability, between, say, LL verb formation operations and UL denominalization operations. This however, is not possible in the case

[10] See Grestenberger and Kastner 2022 for a relevant discussion of semantic predictability in the context of Hebrew denominalization.

of Hebrew because assigning a vector-set representation of the relevant LL operations requires a representation of the root itself, which is not straightforwardly computable by word embedding models, given that roots never appear without a template in corpora.

In English, inputs to both LL and UL operations appear in corpora, and therefore, the case study of English suffixation presented in this paper avoids the pitfalls of the Hebrew data set. We therefore take this case study to be a more robust piece of evidence for the sensitivity of word-embedding models to the two-level distinction.

5. Conclusion

In this study, we provide evidence that word-embedding models are sensitive to the level at which morphological processes take place. We do that by comparing the effect on a word's representation in different word-embedding models of different pairs of morphological operations – each consisting of a lower level (LL) operation and an upper level (UL) one. We test pairs of English suffixes – /-ity/ vs. /-ness/ and /-al/ vs. /-less/ – and show that the effect of the UL ones is more stable (namely, predictable) than that of the LL ones. This expands on our attempt to illustrate the sensitivity of word-embeddings to the level-distinction in Benbaji et al. 2022, which, as discussed above, is insufficient.

Crucially, the models we examine in this paper only have access to, and thus base their vectoral representations of words on, the distribution of each word in the given corpora. Their observed sensitivity to morphological distinctions must therefore be mediated by the semantic effects of these distinctions, which may often seem very subtle. This work, then, serves as an addition to the growing body of evidence indicating that word-embedding models indeed capture fine-grained aspects of word meanings. More broadly this work can be viewed as lending support to the *distributional hypothesis*, according to which the semantics of a word is manifested by its distribution.

References

Allen, Carl, and Timothy Hospedales. 2019. Analogies Explained: Towards Understanding Word Embeddings. In *Proceedings of the 36th International Conference on Machine Learning*, ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov, volume 97 of *Proceedings of Machine Learning Research*, 223–231. PMLR. URL https://proceedings.mlr.press/v97/allen19a.html.

Arad, Maya. 2003. Locality Constraints on the Interpretation of Roots: The Case of Hebrew Denominal Verbs. *Natural Language and Linguistic Theory* 21:737–778. URL https://doi.org/10.1023/a:1025533719905.

Aronoff, Mark. 1976. *Word Formation in Generative Grammar*. Linguistic Inquiry monographs. MIT press.

Aronoff, Mark, and Mark Lindsay. 2014. Productivity, Blocking, and Lexicalization. In *The Oxford Handbook of Derivational Morphology*. Oxford University Press. URL https://doi.org/10.1093/oxfordhb/9780199641642.013.0005.

Benbaji, Ido, Omri Doron, and Adèle Hénot-Mortier. 2022. Word-Embeddings Distinguish Denominal and Root-Derived Verbs in Semitic. *arXiv preprint* URL https://arxiv.org/abs/2208.05721.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint* URL https://arxiv.org/abs/1607.04606.

Bonami, Olivier, and Matías Guzmán Naranjo. 2023. *Distributional evidence for derivational paradigms*, 219–258. Berlin, Boston: De Gruyter. URL https://doi.org/10.1515/9783111074917-008.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English.* New York: Harper & Row.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805.

Ethayarajh, Kawin, David Duvenaud, and Graeme Hirst. 2019. Towards Understanding Linear Word Analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3253–3262. Florence, Italy: Association for Computational Linguistics. URL `https://aclanthology.org/P19-1315`.

Firth, John. 1957. A Synopsis of Linguistic Theory, 1930-1955. *Studies in linguistic analysis* 10–32.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomás Mikolov. 2018. Learning Word Vectors for 157 Languages. *CoRR* abs/1802.06893.

Grestenberger, Laura, and Itamar Kastner. 2022. Directionality in cross-categorial derivations. *Glossa: a journal of general linguistics* 7:1–64.

Harris, Zellig S. 1954. Distributional Structure. *WORD* 10:146–162. URL `https://doi.org/10.1080/00437956.1954.11659520`.

Hénot-Mortier, Adèle. 2022. Evidence for an encoding of morphological blocking effects within two English word embedding models. In *Architectures and Mechanisms of Language Processing 2022 (AMLaP 28)*.

Itai, Alon, and Shuly Wintner. 2019. The Knesset Meetings Corpus 2004-2005. URL `https://doi.org/10.5281/zenodo.2707356`.

Jurafsky, D., and J.H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in artificial intelligence. Pearson Prentice Hall.

Kiparsky, Paul. 1982. Word-formation and the lexicon. In *Papers of the Mid-America Linguistics Conference*.

Lebret, Rémi, and Ronan Collobert. 2013. Word Emdeddings through Hellinger PCA URL `https://arxiv.org/abs/1312.5542`.

Marelli, Marco, and Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review* 122:485–515. URL `https://doi.org/10.1037/a0039267`.

Mikolov, Tomás, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations (ICLR)*, ed. by Yoshua Bengio and Yann LeCun. Scottsdale, Arizona, USA. URL `http://arxiv.org/abs/1301.3781`.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. URL `https://arxiv.org/abs/1310.4546`.

Naranjo, Matías Guzmán, and Olivier Bonami. 2023. A distributional assessment of rivalry in word formation. *Word Structure* 16:87–114. URL `https://doi.org/10.3366/word.2023.0222`.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha, Qatar: Association for Computational Linguistics. URL `https://aclanthology.org/D14-1162`.

Rasin, Ezer, Omer Preminger, and David Pesetsky. 2021. A re-evaluation of Arad's argument for roots URL `https://ling.auf.net/lingbuzz/006077`, available on Lingbuzz.

Seker, Amit, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, and Reut Tsarfaty. 2021. AlephBERT: a Hebrew Large pre-trained Language Model to start-off your Hebrew NLP application with. *CoRR* abs/2104.04052.

Siegel, Dorothy Carla. 1974. Topics in English morphology. Doctoral dissertation, Massachusetts Institute of Technology.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762.

Xiao, Han. 2018. BERT-as-service. `https://github.com/hanxiao/bert-as-service`.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. URL `https://arxiv.org/abs/1904.03310`.

Ido Benbaji-Elhadad, Omri Doron, Adèle Hénot-Mortier
ibenbaji@mit.edu, omrid@mit.edu, mortier@mit.edu