Distinguishing levels of morphological derivations in word-embedding models

Ido Benbaji [1] Omri Doron [1] Adèle Hénot-Mortier [1]

[1]Massachusetts Institute of Technology

August 16, 2022

Introduction: the 2-level model of morphology, and word embeddings

A few basic principles of word formation

The two-level model ([6, 10] a.o.)

- Morphological operations can be of two types...
  - **Lower Level**: idiosyncratic, non-compositional, unpredictable.
  - **Upper Level**: deterministic, compositional, predicatble.
- Given a base element A and a word derived from it B, the two-level hypothesis predicts that both the semantic and the phonological relation between A and B depends on the level at which the derivation takes place.
- While the meaning and form of words that diverge at UL are predicted to be regularly connected, the connection between words that diverge at LL is predicted to be looser.

A few basic principles of word formation

The two-level model ([6, 10] a.o.)

- Morphological operations can be of two types...
  - **Lower Level**: idiosyncratic, non-compositional, unpredictable.
  - **Upper Level**: deterministic, compositional, predicatble.
- Given a base element A and a word derived from it B, the two-level hypothesis predicts that both the semantic and the phonological relation between A and B depends on the level at which the derivation takes place.
- While the meaning and form of words that diverge at UL are predicted to be regularly connected, the connection between words that diverge at LL is predicted to be looser.

A few basic principles of word formation

The two-level model ([6, 10] a.o.)

- Morphological operations can be of two types...
  - **Lower Level**: idiosyncratic, non-compositional, unpredictable.
  - **Upper Level**: deterministic, compositional, predicatble.
- Given a base element A and a word derived from it B, the two-level hypothesis predicts that both the semantic and the phonological relation between A and B depends on the level at which the derivation takes place.
- While the meaning and form of words that diverge at UL are predicted to be regularly connected, the connection between words that diverge at LL is predicted to be looser.

A few basic principles of word formation

The two-level model ([6, 10] a.o.)

- Morphological operations can be of two types...
  - **Lower Level**: idiosyncratic, non-compositional, unpredictable.
  - **Upper Level**: deterministic, compositional, predicatble.
- Given a base element A and a word derived from it B, the two-level hypothesis predicts that both the semantic and the phonological relation between A and B depends on the level at which the derivation takes place.
- While the meaning and form of words that diverge at UL are predicted to be regularly connected, the connection between words that diverge at LL is predicted to be looser.

A few basic principles of word formation

The two-level model ([6, 10] a.o.)

- Morphological operations can be of two types...
  - **Lower Level**: idiosyncratic, non-compositional, unpredictable.
  - **Upper Level**: deterministic, compositional, predicatble.
- Given a base element A and a word derived from it B, the two-level hypothesis predicts that both the semantic and the phonological relation between A and B depends on the level at which the derivation takes place.
- While the meaning and form of words that diverge at UL are predicted to be regularly connected, the connection between words that diverge at LL is predicted to be looser.

Key semantic predictions of the two-level model

We focus on the semantic effects of the level-distinction, which makes two key predictions:

A. **Words derived from the same element via LL operations may arbitrarily differ semantically.**

B. **Words derived from the same element *via* UL operations should be closely related semantically.**

Key semantic predictions of the two-level model

We focus on the semantic effects of the level-distinction, which makes two key predictions:

A. **Words derived from the same element via LL operations may arbitrarily differ semantically.**

B. **Words derived from the same element *via* UL operations should be closely related semantically.**

Key semantic predictions of the two-level model

We focus on the semantic effects of the level-distinction, which makes two key predictions:

A. **Words derived from the same element via LL operations may arbitrarily differ semantically.**
B. **Words derived from the same element *via* UL operations should be closely related semantically.**

What are word embedding models?

- **Word embeddings are high-dimensional vector representations of words**, based their co-occurrence with other words in a corpus. [7].
- They can be "static" (1 word = 1 fixed vector) or "contextualized" (1 word = 1 context-dependent vector).
- Static embeddings include Word2Vec [12], GloVe [13], and fastText [2]; contextualized ones include BERT [3].

What are word embedding models?

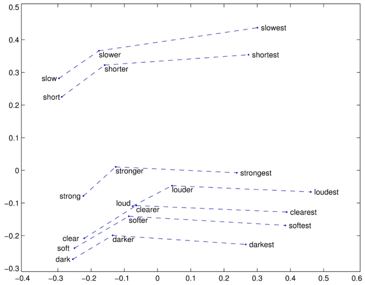
- **Word embeddings are high-dimensional vector representations of words**, based their co-occurrence with other words in a corpus. [7].
- They can be "static" (1 word = 1 fixed vector) or "contextualized" (1 word = 1 context-dependent vector).
- Static embeddings include Word2Vec [12], GloVe [13], and fastText [2]; contextualized ones include BERT [3].

What are word embedding models?

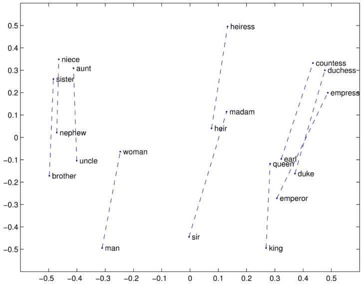
- **Word embeddings are high-dimensional vector representations of words**, based their co-occurrence with other words in a corpus. [7].
- They can be "static" (1 word = 1 fixed vector) or "contextualized" (1 word = 1 context-dependent vector).
- Static embeddings include Word2Vec [12], GloVe [13], and fastText [2]; contextualized ones include BERT [3].

Relevance of word embeddings to our task

- Embeddings come with a robust measure of semantic similarity: **cosine similarity** (∼angle between 2 vectors).
- Past empirical evidence in favor of embeddings' encoding of semantic features and relationships [13].



(a) Positive form → comparative → superlative transformations

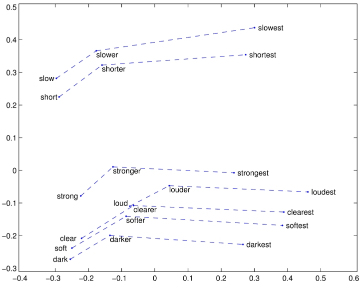


(b) Masculine ↔ feminine transformations

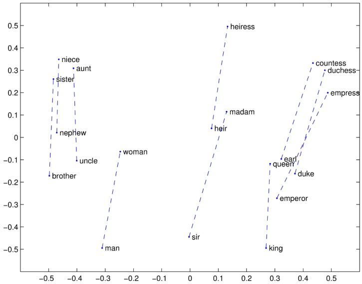
Figure 1: Plots from the original GloVe model [13]

Relevance of word embeddings to our task

- Embeddings come with a robust measure of semantic similarity: **cosine similarity** (∼angle between 2 vectors).
- **Past empirical evidence in favor of embeddings' encoding of semantic features and relationships** [13].



(a) Positive form → comparative → superlative transformations



(b) Masculine ↔ feminine transformations

Figure 1: Plots from the original GloVe model [13]

Case study #1: Hebrew denominal verbs

A bird's eye view on templatic morphology

A non-concatenative system

- **In Modern Hebrew, functional heads are instantiated by "templates".**
- Templates are discontinuous sequences of phonemes (usually vowels), which are intended to be "filled" by root ($\sqrt{}$) consonants.

An illustration of templatic morphology [1]

- For instance, template taCCiC (=*n*-head) can combine with root $\sqrt{\mathrm{x}\int\mathrm{v}}$ to form the word (noun) $\mathrm{tax}\int\mathrm{iv}$, 'calculation'.
- In the above template, the t is called a *templatic consonant*.
- A root, applied to different templates, yields words with very different meanings: $\sqrt{\mathrm{x}\int\mathrm{v}}$+CaCuC=$\mathrm{xa}\int\mathrm{uv}$, 'important', no obvious link with 'calculation'! **In line with prediction A**.

A bird's eye view on templatic morphology

A non-concatenative system

- **In Modern Hebrew, functional heads are instantiated by "templates".**
- Templates are discontinuous sequences of phonemes (usually vowels), which are intended to be "filled" by root ($\sqrt{}$) consonants.

An illustration of templatic morphology [1]

- For instance, template taCCiC (=*n*-head) can combine with root $\sqrt{\mathrm{x}\int\mathrm{v}}$ to form the word (noun) $\mathrm{tax}\int\mathrm{iv}$, 'calculation'.
- In the above template, the t is called a *templatic consonant*.
- A root, applied to different templates, yields words with very different meanings: $\sqrt{\mathrm{x}\int\mathrm{v}}$+CaCuC=$\mathrm{xa}\int\mathrm{uv}$, 'important', no obvious link with 'calculation'! **In line with prediction A**.

A bird's eye view on templatic morphology

A non-concatenative system

- **In Modern Hebrew, functional heads are instantiated by "templates"**.
- Templates are discontinuous sequences of phonemes (usually vowels), which are intended to be "filled" by root ($\sqrt{}$) consonants.

An illustration of templatic morphology [1]

- For instance, template taCCiC (=*n*-head) can combine with root $\sqrt{\mathrm{x}\int\mathrm{v}}$ to form the word (noun) $\mathrm{tax}\int\mathrm{iv}$, 'calculation'.
- In the above template, the t is called a *templatic consonant*.
- A root, applied to different templates, yields words with very different meanings: $\sqrt{\mathrm{x}\int\mathrm{v}}$+CaCuC=$\mathrm{xa}\int\mathrm{uv}$, 'important', no obvious link with 'calculation'! **In line with prediction A**.

A bird's eye view on templatic morphology

A non-concatenative system

- **In Modern Hebrew, functional heads are instantiated by "templates"**.
- Templates are discontinuous sequences of phonemes (usually vowels), which are intended to be "filled" by root ($\surd$) consonants.

An illustration of templatic morphology [1]

- For instance, template taCCiC (=*n*-head) can combine with root $\sqrt{\mathrm{x\int v}}$ to form the word (noun) $\mathrm{tax\int iv}$, 'calculation'.
- In the above template, the t is called a *templatic consonant*.
- A root, applied to different templates, yields words with very different meanings: $\sqrt{\mathrm{x\int v}}$+CaCuC=$\mathrm{xa\int uv}$, 'important', no obvious link with 'calculation'! **In line with prediction A**.

The 2-level model at work in Modern Hebrew

Hebrew denominal verbs

- **Denominal verbs are derived from a noun**. In other words, they result from the merger of a *n*-head (LL), followed by that of a *v*-head (UL).
- It is not easy to tease apart denominals from "basic" verbs derived directly from a root in English corpora (but see [8]).
- **Hebrew comes with a clear diagnostic: templatic consonants**! If a verb contains a consonant that (1) belongs to a known nominal template, and (2) does not belong to the original root; then the verb is probably denominal [1].

$\sqrt{\text{xʃv}}$ → maCCeC (*n*) → maxʃev 'computer' – CiCCeC (*v*) → mixʃev 'computerized' (denominal)

$\sqrt{\text{xʃv}}$ → CiCCeC (*v*) → xiʃev 'calculated'

The 2-level model at work in Modern Hebrew

Hebrew denominal verbs

- **Denominal verbs are derived from a noun**. In other words, they result from the merger of a *n*-head (LL), followed by that of a *v*-head (UL).
- It is not easy to tease apart denominals from "basic" verbs derived directly from a root in English corpora (but see [8]).
- **Hebrew comes with a clear diagnostic: templatic consonants**! If a verb contains a consonant that (1) belongs to a known nominal template, and (2) does not belong to the original root; then the verb is probably denominal [1].

$\sqrt{\text{x}\int\text{v}}$ → maCCeC (*n*) → maxʃev 'computer' – CiCCeC (*v*) → mixʃev 'computerized' (denominal)

$\sqrt{\text{x}\int\text{v}}$ → CiCCeC (*v*) → xiʃev 'calculated'

The 2-level model at work in Modern Hebrew

Hebrew denominal verbs

- **Denominal verbs are derived from a noun**. In other words, they result from the merger of a *n*-head (LL), followed by that of a *v*-head (UL).
- It is not easy to tease apart denominals from "basic" verbs derived directly from a root in English corpora (but see [8]).
- **Hebrew comes with a clear diagnostic: templatic consonants**! If a verb contains a consonant that (1) belongs to a known nominal template, and (2) does not belong to the original root; then the verb is probably denominal [1].

$\sqrt{\text{x}\int\text{v}}$ — maCCeC (*n*) → maxʃev 'computer' – CiCCeC (*v*) → mixʃev 'computerized' (denominal)

$\sqrt{\text{x}\int\text{v}}$ — CiCCeC (*v*) → xiʃev 'calculated'

Denominal *vs* root-derived verbs [1]

- Back to the predictions of the 2-level model...
  - A. If a noun N and a verb V derive from the same *root* (*via* a LL operation), we expect them to **differ semantically** in a somewhat arbitrary way.
  - B. If a denominal verb D derives from a base noun N (*via* a UL operation), we expect them to be **close semantically**.
- Thus, given a root $\sqrt{}$, a noun N, a verb V, a denominal D, s.t. $\sqrt{} \xrightarrow{LL} N$, $\sqrt{} \xrightarrow{LL} V$, and $N \xrightarrow{UL} D$, we expect:

$$\mathcal{S}(N, D) > \mathcal{S}(N, V)$$

For some well-chosen semantic measure $\mathcal{S}$ between pairs of words. Building on the previous example:

$$\mathcal{S}(\mathrm{max}\int\mathrm{ev}_N, \mathrm{mix}\int\mathrm{ev}_D) > \mathcal{S}(\mathrm{max}\int\mathrm{ev}_N, \mathrm{xi}\int\mathrm{ev}_V)$$

Denominal *vs* root-derived verbs [1]

- Back to the predictions of the 2-level model...
  A. If a noun N and a verb V derive from the same *root* (*via* a LL operation), we expect them to **differ semantically** in a somewhat arbitrary way.
  B. If a denominal verb D derives from a base noun N (*via* a UL operation), we expect them to be **close semantically**.
- Thus, given a root $\sqrt{}$, a noun N, a verb V, a denominal D, s.t. $\sqrt{} \xrightarrow{LL} N$, $\sqrt{} \xrightarrow{LL} V$, and $N \xrightarrow{UL} D$, we expect:

$$\mathcal{S}(N, D) > \mathcal{S}(N, V)$$

For some well-chosen semantic measure $\mathcal{S}$ between pairs of words. Building on the previous example:

$$\mathcal{S}(\mathrm{max\int ev}_N, \mathrm{mix\int ev}_D) > \mathcal{S}(\mathrm{max\int ev}_N, \mathrm{xi\int ev}_V)$$

Denominal *vs* root-derived verbs [1]

- Back to the predictions of the 2-level model...
  A. If a noun N and a verb V derive from the same *root* (*via* a LL operation), we expect them to **differ semantically** in a somewhat arbitrary way.
  B. If a denominal verb D derives from a base noun N (*via* a UL operation), we expect them to be **close semantically**.
- Thus, given a root $\sqrt{}$, a noun N, a verb V, a denominal D, s.t. $\sqrt{} \xrightarrow{LL} N$, $\sqrt{} \xrightarrow{LL} V$, and $N \xrightarrow{UL} D$, we expect:

$$\mathcal{S}(N, D) > \mathcal{S}(N, V)$$

For some well-chosen semantic measure $\mathcal{S}$ between pairs of words. Building on the previous example:

$$\mathcal{S}(\mathrm{max{\int}ev}_N, \mathrm{mix{\int}ev}_D) > \mathcal{S}(\mathrm{max{\int}ev}_N, \mathrm{xi{\int}ev}_V)$$

Denominal *vs* root-derived verbs [1]

- Back to the predictions of the 2-level model...
  A. If a noun N and a verb V derive from the same *root* (*via* a LL operation), we expect them to **differ semantically** in a somewhat arbitrary way.
  B. If a denominal verb D derives from a base noun N (*via* a UL operation), we expect them to be **close semantically**.
- Thus, given a root $\sqrt{}$, a noun N, a verb V, a denominal D, s.t. $\sqrt{} \xrightarrow{LL} N$, $\sqrt{} \xrightarrow{LL} V$, and $N \xrightarrow{UL} D$, we expect:

$$\mathcal{S}(N, D) > \mathcal{S}(N, V)$$

For some well-chosen semantic measure $\mathcal{S}$ between pairs of words. Building on the previous example:

$$\mathcal{S}(\mathrm{max}\!\int\!\mathrm{ev}_N, \mathrm{mix}\!\int\!\mathrm{ev}_D) > \mathcal{S}(\mathrm{max}\!\int\!\mathrm{ev}_N, \mathrm{xi}\!\int\!\mathrm{ev}_V)$$

Denominal *vs* root-derived verbs [1]

- Back to the predictions of the 2-level model...
  A. If a noun N and a verb V derive from the same *root* (*via* a LL operation), we expect them to **differ semantically** in a somewhat arbitrary way.
  B. If a denominal verb D derives from a base noun N (*via* a UL operation), we expect them to be **close semantically**.
- Thus, given a root $\sqrt{}$, a noun N, a verb V, a denominal D, s.t. $\sqrt{} \xrightarrow{LL} N$, $\sqrt{} \xrightarrow{LL} V$, and $N \xrightarrow{UL} D$, we expect:

$$\mathcal{S}(N, D) > \mathcal{S}(N, V)$$

For some well-chosen semantic measure $\mathcal{S}$ between pairs of words. Building on the previous example:

$$\mathcal{S}(\mathrm{max{\int}ev}_N, \mathrm{mix{\int}ev}_D) > \mathcal{S}(\mathrm{max{\int}ev}_N, \mathrm{xi{\int}ev}_V)$$

How does the 2-level model translate into a word embedding?

- Let us define $Area(\sqrt{})$ as the subspace (convex envelope?) of $\{\overrightarrow{X} | \sqrt{} \rightarrow^* X\}$. The predictions of the 2-level model become:
  - A. Given a root $\sqrt{}$, and 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 very close to each other within $Area(\sqrt{})$.
- Let $\sqrt{}$, N, D, $(V_i)_{i \in [1,K]}$, be s.t. $\sqrt{} \xrightarrow{LL} N$, $\forall i \in [1, K]$ $\sqrt{} \xrightarrow{LL} V_i$, and $N \xrightarrow{UL} D$. We predict:

$$CosSim(\overrightarrow{N}, \overrightarrow{D}) > \max_i CosSim(\overrightarrow{N}, \overrightarrow{V_i}) \quad \text{(Stronger Hypothesis[1])}$$

$$CosSim(\overrightarrow{N}, \overrightarrow{D}) > \frac{1}{K} \sum_{i=1}^{K} CosSim(\overrightarrow{N}, \overrightarrow{V_i}) \quad \text{(Weaker Hypothesis)}$$

[1] The stronger hypothesis is not expected to hold all the time, because the closest $\overrightarrow{V_i}$ may accidentally end up closer to $\overrightarrow{N}$ than $\overrightarrow{D}$ is, due to the arbitrariness of LL operations. This motivates the use of the weaker hypothesis.

How does the 2-level model translate into a word embedding?

- Let us define $Area(\sqrt{})$ as the subspace (convex envelope?) of $\{\overrightarrow{X} | \sqrt{} \rightarrow^* X\}$. The predictions of the 2-level model become:
  - A. Given a root $\sqrt{}$, and 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 very close to each other within $Area(\sqrt{})$.
- Let $\sqrt{}$, N, D, $(V_i)_{i \in [1,K]}$, be s.t. $\sqrt{} \xrightarrow{LL} N$, $\forall i \in [1, K]$ $\sqrt{} \xrightarrow{LL} V_i$, and $N \xrightarrow{UL} D$. We predict:

$$CosSim(\overrightarrow{N}, \overrightarrow{D}) > \max_i CosSim(\overrightarrow{N}, \overrightarrow{V_i}) \quad \text{(Stronger Hypothesis[1])}$$

$$CosSim(\overrightarrow{N}, \overrightarrow{D}) > \frac{1}{K} \sum_{i=1}^{K} CosSim(\overrightarrow{N}, \overrightarrow{V_i}) \quad \text{(Weaker Hypothesis)}$$

[1] The stronger hypothesis is not expected to hold all the time, because the closest $\overrightarrow{V_i}$ may accidentally end up closer to $\overrightarrow{N}$ than $\overrightarrow{D}$ is, due to the arbitrariness of LL operations. This motivates the use of the weaker hypothesis.

How does the 2-level model translate into a word embedding?

- Let us define $Area(\sqrt{})$ as the subspace (convex envelope?) of $\{\overrightarrow{X} | \sqrt{} \rightarrow^* X\}$. The predictions of the 2-level model become:
    - A. Given a root $\sqrt{}$, and 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 very close to each other within $Area(\sqrt{})$.
- Let $\sqrt{}$, N, D, $(V_i)_{i \in [1,K]}$, be s.t. $\sqrt{} \xrightarrow{LL} N$, $\forall i \in [1, K]$ $\sqrt{} \xrightarrow{LL} V_i$, and $N \xrightarrow{UL} D$. We predict:

$$CosSim(\overrightarrow{N}, \overrightarrow{D}) > \max_i CosSim(\overrightarrow{N}, \overrightarrow{V_i}) \quad \text{(Stronger Hypothesis[1])}$$

$$CosSim(\overrightarrow{N}, \overrightarrow{D}) > \frac{1}{K} \sum_{i=1}^{K} CosSim(\overrightarrow{N}, \overrightarrow{V_i}) \quad \text{(Weaker Hypothesis)}$$

[1] The stronger hypothesis is not expected to hold all the time, because the closest $\overrightarrow{V_i}$ may accidentally end up closer to $\overrightarrow{N}$ than $\overrightarrow{D}$ is, due to the arbitrariness of LL operations. This motivates the use of the weaker hypothesis.

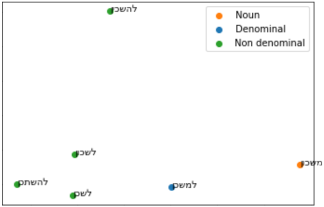
How does the 2-level model translate into a word embedding?

- Let us define $Area(\surd)$ as the subspace (convex envelope?) of $\{\overrightarrow{X} | \surd \rightarrow^* X\}$. The predictions of the 2-level model become:
  - A. Given a root $\surd$, and A, B, s.t. $\surd \overset{LL}{\rightarrow} A$, and $\surd \overset{LL}{\rightarrow} B$, we expect $\overrightarrow{A}$ and $\overrightarrow{B}$ to be randomly distributed across $Area(\surd)$.
  - B. Given $\surd$, A and B, s.t. $\surd \overset{LL}{\rightarrow} A \overset{UL}{\rightarrow} B$, we expect $\overrightarrow{A}$ and $\overrightarrow{B}$ to be very close to each other within $Area(\surd)$.
- Let $\surd$, N, D, $(V_i)_{i \in [1,K]}$, be s.t. $\surd \overset{LL}{\rightarrow} N$, $\forall i \in [1, K]$ $\surd \overset{LL}{\rightarrow} V_i$, and $N \overset{UL}{\rightarrow} D$. We predict:

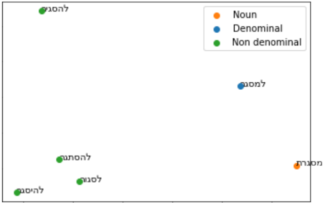
$$CosSim(\overrightarrow{N}, \overrightarrow{D}) > \max_i CosSim(\overrightarrow{N}, \overrightarrow{V_i}) \quad \text{(Stronger Hypothesis}^1\text{)}$$

$$CosSim(\overrightarrow{N}, \overrightarrow{D}) > \frac{1}{K} \sum_{i=1}^{K} CosSim(\overrightarrow{N}, \overrightarrow{V_i}) \quad \text{(Weaker Hypothesis)}$$

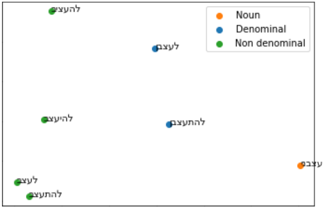
[1] The stronger hypothesis is not expected to hold all the time, because the closest $\overrightarrow{V_i}$ may accidentally end up closer to $\overrightarrow{N}$ than $\overrightarrow{D}$ is, due to the arbitrariness of LL operations. This motivates the use of the weaker hypothesis.



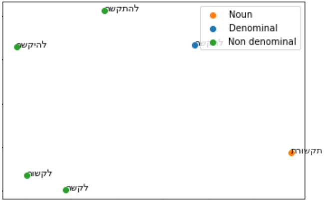
(a) Noun: ‘pawning’; Denominal: ‘to pawn’



(b) Noun: ‘frame’; Denominal: ‘to frame’



(c) Noun: ‘annoyed’; Denominals: ‘to get annoyed’, ‘to annoy’



(d) Noun: ‘communication’; Denominal: ‘to communicate’

Figure 2: 2D-reduction of a few datapoints (PCA, cosine kernel, fastText)

Results for the Hebrew Case study

- We tested 4 architectures: Word2vec [12], GloVe [13], fastText [4], BERT [14]. The last 2 were pretrained.
- **Weaker hypothesis** ($CosSim(\overrightarrow{N}, \overrightarrow{D}) / \frac{1}{K} \sum_{i=1}^{K} CosSim(\overrightarrow{N}, \overrightarrow{V_i})$):
  - All Wilcoxon tests appear significant.
  - Large effect sizes, except for BERT.
- **Stronger hypothesis** ($CosSim(\overrightarrow{N}, \overrightarrow{D}) / \max_i CosSim(\overrightarrow{N}, \overrightarrow{V_i})$):
  - All Wilcoxon tests but two ($GloVe_{50}$, BERT) are significant.
  - Large effect sizes on the significant results, except for $GloVe_{100}$.

	$Word2Vec_{100}$	$GloVe_{50}$	$GloVe_{100}$	$fastText_{300}$	$BERT_{768}$
# datapoints	31	31	31	53	66
Weak hyp. (mean)	1e-6 .86 (L)	2e-4 .52 (L)	7e-5 .66 (L)	1e-10 .79 (L)	5e-4 .30 (S)
Strong hyp. (max)	4e-5 .66 (L)	2e-1 .06 (N)	3e-2 .20 (S)	1e-8 .62 (L)	4e-1 .02 (N)

Table 1: *p*-values (1-tailed Wilcoxon) and effect sizes (Cliff's Δ; N=Negligible; S=Small; M=Medium; L=Large) for the weak and strong hypotheses and 5 embedding models

Results for the Hebrew Case study

- We tested 4 architectures: Word2vec [12], GloVe [13], fastText [4], BERT [14]. The last 2 were pretrained.
- **Weaker hypothesis** ($CosSim(\overrightarrow{N}, \overrightarrow{D}) / \frac{1}{K} \sum_{i=1}^{K} CosSim(\overrightarrow{N}, \overrightarrow{V_i})$):
  - All Wilcoxon tests appear significant.
  - Large effect sizes, except for BERT.
- **Stronger hypothesis** ($CosSim(\overrightarrow{N}, \overrightarrow{D}) / \max_i CosSim(\overrightarrow{N}, \overrightarrow{V_i})$):
  - All Wilcoxon tests but two ($GloVe_{50}$, BERT) are significant.
  - Large effect sizes on the significant results, except for $GloVe_{100}$.

	$Word2Vec_{100}$	$GloVe_{50}$	$GloVe_{100}$	$fastText_{300}$	$BERT_{768}$
# datapoints	31	31	31	53	66
Weak hyp. (mean)	1e-6 .86 (L)	2e-4 .52 (L)	7e-5 .66 (L)	1e-10 .79 (L)	5e-4 .30 (S)
Strong hyp. (max)	4e-5 .66 (L)	2e-1 .06 (N)	3e-2 .20 (S)	1e-8 .62 (L)	4e-1 .02 (N)

Table 1: p-values (1-tailed Wilcoxon) and effect sizes (Cliff's Δ; N=Negligible; S=Small; M=Medium; L=Large) for the weak and strong hypotheses and 5 embedding models

Results for the Hebrew Case study

- We tested 4 architectures: Word2vec [12], GloVe [13], fastText [4], BERT [14]. The last 2 were pretrained.
- **Weaker hypothesis** $(CosSim(\vec{N}, \vec{D}) / \frac{1}{K} \sum_{i=1}^{K} CosSim(\vec{N}, \vec{V_i}))$:
  - All Wilcoxon tests appear significant.
  - Large effect sizes, except for BERT.
- **Stronger hypothesis** $(CosSim(\vec{N}, \vec{D}) / \max_i CosSim(\vec{N}, \vec{V_i}))$:
  - All Wilcoxon tests but two (GloVe_{50}, BERT) are significant.
  - Large effect sizes on the significant results, except for GloVe_{100}.

	Word2Vec_{100}	GloVe_{50}	GloVe_{100}	fastText_{300}	BERT_{768}
# datapoints	31	31	31	53	66
Weak hyp. (mean)	1e-6 .86 (L)	2e-4 .52 (L)	7e-5 .66 (L)	1e-10 .79 (L)	5e-4 .30 (S)
Strong hyp. (max)	4e-5 .66 (L)	2e-1 .06 (N)	3e-2 .20 (S)	1e-8 .62 (L)	4e-1 .02 (N)

Table 1: p-values (1-tailed Wilcoxon) and effect sizes (Cliff's Δ; N=Negligible; S=Small; M=Medium; L=Large) for the weak and strong hypotheses and 5 embedding models

Case study #2: English suffixation

English suffixation and stress

- English suffixes that can apparently attach to the same kind of base have **different effects on stress assignment** [9].
- On "adjective-like" bases for instance, ***-ity* shifts stress while *-ness* doesn't** (*glóbal* → *globálity*, *glóbalness*).
  - Means that *-ity* has access to the phonological features of its base, while *-ness* doesn't (phonological opacity)...
  - Suggests that *-ity* attaches to an uncategorized *root* to form a noun and participates in a LL operation, while *-ness* attaches to a *word* (adjective) and participates in an UL operation.

Predictions regarding the semantic effect of *-ity* and *-ness*

- Assuming phonological opacity correlates with "semantic" opacity, ***-ity*-affixation (LL) should yield more variable meanings on average than *-ness*-affixation (UL)**.
- The prediction can extend to other LL/UL pairs of suffixes, like *-al*/*-less* (see Appendix II for results).

English suffixation and stress

- English suffixes that can apparently attach to the same kind of base have **different effects on stress assignment** [9].
- On "adjective-like" bases for instance, ***-ity* shifts stress while *-ness* doesn't** (*glóbal* → *globálity*, *glóbalness*).
  - Means that *-ity* has access to the phonological features of its base, while *-ness* doesn't (phonological opacity)...
  - Suggests that *-ity* attaches to an uncategorized *root* to form a noun and participates in a LL operation, while *-ness* attaches to a *word* (adjective) and participates in an UL operation.

Predictions regarding the semantic effect of *-ity* and *-ness*

- Assuming phonological opacity correlates with "semantic" opacity, ***-ity*-affixation (LL) should yield more variable meanings on average than *-ness*-affixation (UL)**.
- The prediction can extend to other LL/UL pairs of suffixes, like *-al*/*-less* (see Appendix II for results).

English suffixation and stress

- English suffixes that can apparently attach to the same kind of base have **different effects on stress assignment** [9].
- On "adjective-like" bases for instance, ***-ity* shifts stress while *-ness* doesn't** (*glóbal* → *globálity*, *glóbalness*).
  - Means that *-ity* has access to the phonological features of its base, while *-ness* doesn't (phonological opacity)...
  - Suggests that *-ity* attaches to an uncategorized *root* to form a noun and participates in a LL operation, while *-ness* attaches to a *word* (adjective) and participates in an UL operation.

Predictions regarding the semantic effect of *-ity* and *-ness*

- Assuming phonological opacity correlates with "semantic" opacity, ***-ity*-affixation (LL) should yield more variable meanings on average than *-ness*-affixation (UL)**.
- The prediction can extend to other LL/UL pairs of suffixes, like *-al*/*-less* (see Appendix II for results).

English suffixation and stress

- English suffixes that can apparently attach to the same kind of base have **different effects on stress assignment** [9].
- On "adjective-like" bases for instance, ***-ity* shifts stress while *-ness* doesn't** (*glóbal* → *globálity*, *glóbalness*).
  - Means that *-ity* has access to the phonological features of its base, while *-ness* doesn't (phonological opacity)...
  - Suggests that *-ity* attaches to an uncategorized *root* to form a noun and participates in a LL operation, while *-ness* attaches to a *word* (adjective) and participates in an UL operation.

Predictions regarding the semantic effect of *-ity* and *-ness*

- Assuming phonological opacity correlates with "semantic" opacity, ***-ity*-affixation (LL) should yield more variable meanings on average than *-ness*-affixation (UL)**.
- The prediction can extend to other LL/UL pairs of suffixes, like *-al*/*-less* (see Appendix II for results).

English suffixation and stress

- English suffixes that can apparently attach to the same kind of base have **different effects on stress assignment** [9].
- On "adjective-like" bases for instance, ***-ity* shifts stress while *-ness* doesn't** (*glóbal* → *globálity*, *glóbalness*).
  - Means that *-ity* has access to the phonological features of its base, while *-ness* doesn't (phonological opacity)...
  - Suggests that *-ity* attaches to an uncategorized *root* to form a noun and participates in a LL operation, while *-ness* attaches to a *word* (adjective) and participates in an UL operation.

Predictions regarding the semantic effect of *-ity* and *-ness*

- Assuming phonological opacity correlates with "semantic" opacity, ***-ity*-affixation (LL) should yield more variable meanings on average than *-ness*-affixation (UL)**.
- The prediction can extend to other LL/UL pairs of suffixes, like *-al*/*-less* (see Appendix II for results).

English suffixation and stress

- English suffixes that can apparently attach to the same kind of base have **different effects on stress assignment** [9].
- On "adjective-like" bases for instance, ***-ity* shifts stress while *-ness* doesn't** (*glóbal* → *globálity*, *glóbalness*).
  - Means that *-ity* has access to the phonological features of its base, while *-ness* doesn't (phonological opacity)...
  - Suggests that *-ity* attaches to an uncategorized *root* to form a noun and participates in a LL operation, while *-ness* attaches to a *word* (adjective) and participates in an UL operation.

Predictions regarding the semantic effect of *-ity* and *-ness*

- Assuming phonological opacity correlates with "semantic" opacity, **-*ity*-affixation (LL) should yield more variable meanings on average than -*ness*-affixation (UL)**.
- The prediction can extend to other LL/UL pairs of suffixes, like *-al*/*-less* (see Appendix II for results).

Modeling the prediction

- For n triplets (a, a-ity, a-$ness$) we compute $\overrightarrow{\textit{-ity}} = \overrightarrow{\textit{a-ity}} - \vec{a}$ and $\overrightarrow{\textit{-ness}} = \overrightarrow{\textit{a-ness}} - \vec{a}$ using embeddings.
- We test if the set of $\overrightarrow{\textit{-ity}}$ vectors exhibits more variability than the set of $\overrightarrow{-ness}$ vectors. Two possible measures:
  - **"Dispersion"**: pairwise CosSim between all the vectors within a set. $\frac{n(n-1)}{2}$ measures per set.
  - **"Variation"**: CosSim between all the vectors of a set and its center (mean vector). n measures per set.

(a) "Dispersion" (b) "Variation"

Figure 3: Measuring variability amongst vectors

Modeling the prediction

- For n triplets (a, a-ity, a-$ness$) we compute $\overrightarrow{\textit{-ity}} = \overrightarrow{\textit{a-ity}} - \overrightarrow{a}$ and $\overrightarrow{\textit{-ness}} = \overrightarrow{\textit{a-ness}} - \overrightarrow{a}$ using embeddings.
- We test if the set of $\overrightarrow{\textit{-ity}}$ vectors exhibits more variability than the set of $\overrightarrow{-\textit{ness}}$ vectors. Two possible measures:
  - **"Dispersion"**: pairwise CosSim between all the vectors within a set. $\frac{n(n-1)}{2}$ measures per set.
  - **"Variation"**: CosSim between all the vectors of a set and its center (mean vector). n measures per set.

(a) "Dispersion" (b) "Variation"

Figure 3: Measuring variability amongst vectors

Modeling the prediction

- For n triplets (a, a-ity, a-$ness$) we compute $\overrightarrow{\textit{-ity}} = \overrightarrow{\textit{a-ity}} - \overrightarrow{a}$ and $\overrightarrow{\textit{-ness}} = \overrightarrow{\textit{a-ness}} - \overrightarrow{a}$ using embeddings.
- We test if the set of $\overrightarrow{\textit{-ity}}$ vectors exhibits more variability than the set of $\overrightarrow{-\textit{ness}}$ vectors. Two possible measures:
  - **"Dispersion"**: pairwise CosSim between all the vectors within a set. $\frac{n(n-1)}{2}$ measures per set.
  - **"Variation"**: CosSim between all the vectors of a set and its center (mean vector). n measures per set.

(a) "Dispersion"   (b) "Variation"

Figure 3: Measuring variability amongst vectors

Modeling the prediction

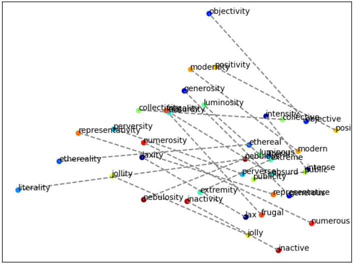
- For n triplets (a, $a\text{-}ity$, $a\text{-}ness$) we compute $\overrightarrow{\text{-}ity} = \overrightarrow{a\text{-}ity} - \vec{a}$ and $\overrightarrow{\text{-}ness} = \overrightarrow{a\text{-}ness} - \vec{a}$ using embeddings.
- We test if the set of $\overrightarrow{\text{-}ity}$ vectors exhibits more variability than the set of $\overrightarrow{-ness}$ vectors. Two possible measures:
  - **"Dispersion"**: pairwise CosSim between all the vectors within a set. $\frac{n(n-1)}{2}$ measures per set.
  - **"Variation"**: CosSim between all the vectors of a set and its center (mean vector). n measures per set.

(a) "Dispersion" (b) "Variation"

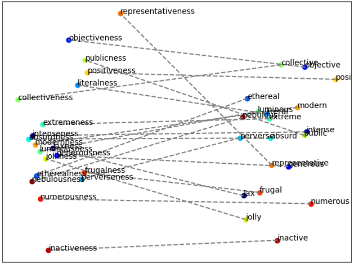
Figure 3: Measuring variability amongst vectors

Characteristics of the embeddings

- We tested the 4 same architectures (Word2Vec [12], GloVe [13], fastText [11], BERT [3]), all pretrained.
  - The first 3 (static) models had an initial dimension of 300;
  - BERT had a dimension of 768 (corresponding to that of its second-to-last layer, used to extract the vectors).



(a) $\overrightarrow{-ity}$ vectors

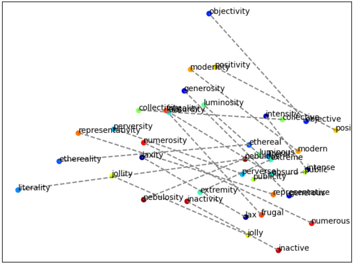


(b) $\overrightarrow{-ness}$ vectors

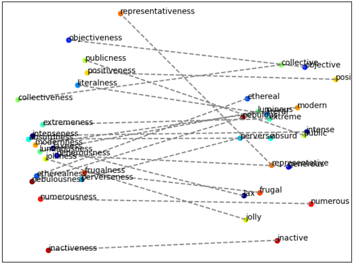
Figure 4: 2D PCA reduction (cosine kernel) of 20 adjective/noun pairs embedded using GloVe_{300}. Lines represent the effect of suffixation.

Characteristics of the embeddings

- We tested the 4 same architectures (Word2Vec [12], GloVe [13], fastText [11], BERT [3]), all pretrained.
  - The first 3 (static) models had an initial dimension of 300;
  - BERT had a dimension of 768 (corresponding to that of its second-to-last layer, used to extract the vectors).



(a) $\overrightarrow{-ity}$ vectors

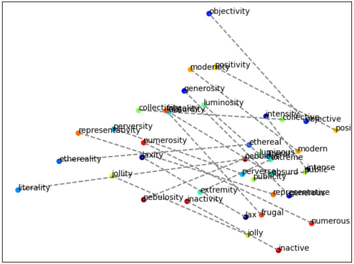


(b) $\overrightarrow{-ness}$ vectors

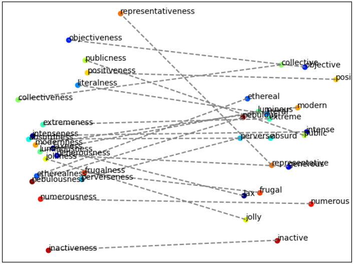
Figure 4: 2D PCA reduction (cosine kernel) of 20 adjective/noun pairs embedded using GloVe$_{300}$. Lines represent the effect of suffixation.

Characteristics of the embeddings

- We tested the 4 same architectures (Word2Vec [12], GloVe [13], fastText [11], BERT [3]), all pretrained.
  - The first 3 (static) models had an initial dimension of 300;
  - BERT had a dimension of 768 (corresponding to that of its second-to-last layer, used to extract the vectors).



(a) $\overrightarrow{-ity}$ vectors

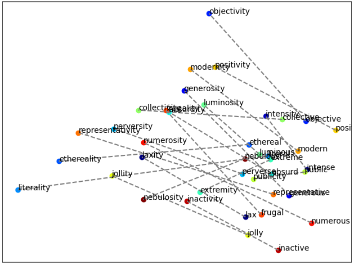


(b) $\overrightarrow{-ness}$ vectors

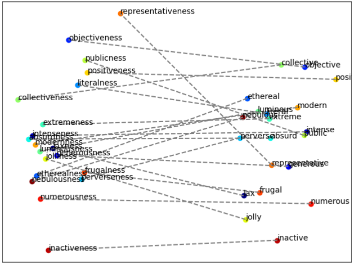
Figure 4: 2D PCA reduction (cosine kernel) of 20 adjective/noun pairs embedded using GloVe$_{300}$. Lines represent the effect of suffixation.

Characteristics of the embeddings

- We tested the 4 same architectures (Word2Vec [12], GloVe [13], fastText [11], BERT [3]), all pretrained.
  - The first 3 (static) models had an initial dimension of 300;
  - BERT had a dimension of 768 (corresponding to that of its second-to-last layer, used to extract the vectors).



(a) $\overrightarrow{-ity}$ vectors



(b) $\overrightarrow{-ness}$ vectors

Figure 4: 2D PCA reduction (cosine kernel) of 20 adjective/noun pairs embedded using GloVe_{300}. Lines represent the effect of suffixation.

Results for the English Case study

- Dispersion contrast significant for all models, small to medium effect sizes.
- Variation contrast significant for all models but Word2Vec (only marginally significant), medium to large effect sizes.
- Confirms that **the semantic effect of -*ness* affixation is less arbitrary than that of -*ity* affixation** in word embeddings.

	Word2Vec	GloVe	fastText	BERT
n	29	126	144	610
"dispersion"	1e-10 .21 (S)	0 .46 (M)	0 .38 (M)	0 .30 (S)
"variation"	.07 .21 (S)	5e-12 .46 (M)	4e-13 .40 (M)	2e-59 .49 (L)

Table 2: *p*-values (2-tailed Wilcoxon) and effect sizes (Cliff's Δ; N=Negligible; S=Small; M=Medium; L=Large) for dispersion and variation measures and 4 embedding models

Results for the English Case study

- Dispersion contrast significant for all models, small to medium effect sizes.
- Variation contrast significant for all models but Word2Vec (only marginally significant), medium to large effect sizes.
- Confirms that **the semantic effect of *-ness* affixation is less arbitrary than that of *-ity* affixation** in word embeddings.

	Word2Vec	GloVe	fastText	BERT
n	29	126	144	610
"dispersion"	1e-10 .21 (S)	0 .46 (M)	0 .38 (M)	0 .30 (S)
"variation"	.07 .21 (S)	5e-12 .46 (M)	4e-13 .40 (M)	2e-59 .49 (L)

Table 2: *p*-values (2-tailed Wilcoxon) and effect sizes (Cliff's Δ; N=Negligible; S=Small; M=Medium; L=Large) for dispersion and variation measures and 4 embedding models

Results for the English Case study

- Dispersion contrast significant for all models, small to medium effect sizes.
- Variation contrast significant for all models but Word2Vec (only marginally significant), medium to large effect sizes.
- Confirms that **the semantic effect of -*ness* affixation is less arbitrary than that of -*ity* affixation** in word embeddings.

	Word2Vec	GloVe	fastText	BERT
n	29	126	144	610
"dispersion"	1e-10 .21 (S)	0 .46 (M)	0 .38 (M)	0 .30 (S)
"variation"	.07 .21 (S)	5e-12 .46 (M)	4e-13 .40 (M)	2e-59 .49 (L)

Table 2: *p*-values (2-tailed Wilcoxon) and effect sizes (Cliff's Δ; N=Negligible; S=Small; M=Medium; L=Large) for dispersion and variation measures and 4 embedding models

Conclusion and Discussion

Conclusion

- We brought evidence in support of **word embeddings' distinguishing between levels of morphological derivation**:
  - In Hebrew, contrast between denominal and root-derived verbs w.r.t. how close they are to the relevant root-derived noun.
  - In English, contrast between pairs of affixes w.r.t. how stable their effect is on the base word.
- We tested a variety of language models, showing that **the prediction was quite robust**.
- Models that did not verify the hypothesis were often tested on **smaller datasets** (e.g. Word2Vec in the English case study); or were characterized by a **fairly small initial dimensionality** (e.g. GloVe_{50} in the Hebrew case study).
- The failure of BERT in the Hebrew study for the stronger hypothesis remains relatively unclear.

Conclusion

- We brought evidence in support of **word embeddings' distinguishing between levels of morphological derivation**:
  - In Hebrew, contrast between denominal and root-derived verbs w.r.t. how close they are to the relevant root-derived noun.
  - In English, contrast between pairs of affixes w.r.t. how stable their effect is on the base word.
- We tested a variety of language models, showing that **the prediction was quite robust**.
- Models that did not verify the hypothesis were often tested on **smaller datasets** (e.g. Word2Vec in the English case study); or were characterized by a **fairly small initial dimensionality** (e.g. GloVe$_{50}$ in the Hebrew case study).
- The failure of BERT in the Hebrew study for the stronger hypothesis remains relatively unclear.

Conclusion

- We brought evidence in support of **word embeddings' distinguishing between levels of morphological derivation**:
  - In Hebrew, contrast between denominal and root-derived verbs w.r.t. how close they are to the relevant root-derived noun.
  - In English, contrast between pairs of affixes w.r.t. how stable their effect is on the base word.
- We tested a variety of language models, showing that **the prediction was quite robust**.
- Models that did not verify the hypothesis were often tested on **smaller datasets** (e.g. Word2Vec in the English case study); or were characterized by a **fairly small initial dimensionality** (e.g. GloVe_{50} in the Hebrew case study).
- The failure of BERT in the Hebrew study for the stronger hypothesis remains relatively unclear.

Conclusion

- We brought evidence in support of **word embeddings' distinguishing between levels of morphological derivation**:
    - In Hebrew, contrast between denominal and root-derived verbs w.r.t. how close they are to the relevant root-derived noun.
    - In English, contrast between pairs of affixes w.r.t. how stable their effect is on the base word.
- We tested a variety of language models, showing that **the prediction was quite robust**.
- Models that did not verify the hypothesis were often tested on **smaller datasets** (e.g. Word2Vec in the English case study); or were characterized by a **fairly small initial dimensionality** (e.g. GloVe_{50} in the Hebrew case study).
- The failure of BERT in the Hebrew study for the stronger hypothesis remains relatively unclear.

Conclusion

- We brought evidence in support of **word embeddings' distinguishing between levels of morphological derivation**:
  - In Hebrew, contrast between denominal and root-derived verbs w.r.t. how close they are to the relevant root-derived noun.
  - In English, contrast between pairs of affixes w.r.t. how stable their effect is on the base word.
- We tested a variety of language models, showing that **the prediction was quite robust**.
- Models that did not verify the hypothesis were often tested on **smaller datasets** (e.g. Word2Vec in the English case study); or were characterized by a **fairly small initial dimensionality** (e.g. GloVe_{50} in the Hebrew case study).
- The failure of BERT in the Hebrew study for the stronger hypothesis remains relatively unclear.

Conclusion

- We brought evidence in support of **word embeddings' distinguishing between levels of morphological derivation**:
  - In Hebrew, contrast between denominal and root-derived verbs w.r.t. how close they are to the relevant root-derived noun.
  - In English, contrast between pairs of affixes w.r.t. how stable their effect is on the base word.
- We tested a variety of language models, showing that **the prediction was quite robust**.
- Models that did not verify the hypothesis were often tested on **smaller datasets** (e.g. Word2Vec in the English case study); or were characterized by a **fairly small initial dimensionality** (e.g. GloVe_{50} in the Hebrew case study).
- The failure of BERT in the Hebrew study for the stronger hypothesis remains relatively unclear.

Conclusion

- We brought evidence in support of **word embeddings' distinguishing between levels of morphological derivation**:
  - In Hebrew, contrast between denominal and root-derived verbs w.r.t. how close they are to the relevant root-derived noun.
  - In English, contrast between pairs of affixes w.r.t. how stable their effect is on the base word.
- We tested a variety of language models, showing that **the prediction was quite robust**.
- Models that did not verify the hypothesis were often tested on **smaller datasets** (e.g. Word2Vec in the English case study); or were characterized by a **fairly small initial dimensionality** (e.g. GloVe_{50} in the Hebrew case study).
- The failure of BERT in the Hebrew study for the stronger hypothesis remains relatively unclear.

Caveats and avenue for future work

- **Written Hebrew, being usually devoid of vowels, is characterized by a high degree of ambiguity!**
- We tried to control for this by using maximally unambiguous forms (e.g. plural). Two potential alternatives:
  - **Use contextual word embeddings to disambiguate**. However, this relocates the issue in the choice of a "suitable" context for each target word.
  - **Train models on textual data including vowels markings** (*niqqud*). This would probably involve *niqqud*-izing existing datasets... with Machine Learning (!)
- Pairs of English suffixes are more or less frequent on a given base... **what if the difference of variability observed for e.g. *-ity* and *-ness* was due to different amounts of noise coming from frequency contrasts?** Appendix II shows some posthoc stats that tend to exclude this eventuality.

Caveats and avenue for future work

- **Written Hebrew, being usually devoid of vowels, is characterized by a high degree of ambiguity!**
- We tried to control for this by using maximally unambiguous forms (e.g. plural). Two potential alternatives:
  - **Use contextual word embeddings to disambiguate.** However, this relocates the issue in the choice of a "suitable" context for each target word.
  - **Train models on textual data including vowels markings** (*niqqud*). This would probably involve *niqqud*-izing existing datasets... with Machine Learning (!)
- Pairs of English suffixes are more or less frequent on a given base... **what if the difference of variability observed for e.g. *-ity* and *-ness* was due to different amounts of noise coming from frequency contrasts?** Appendix II shows some posthoc stats that tend to exclude this eventuality.

Caveats and avenue for future work

- **Written Hebrew, being usually devoid of vowels, is characterized by a high degree of ambiguity!**
- We tried to control for this by using maximally unambiguous forms (e.g. plural). Two potential alternatives:
  - **Use contextual word embeddings to disambiguate.** However, this relocates the issue in the choice of a "suitable" context for each target word.
  - **Train models on textual data including vowels markings** (*niqqud*). This would probably involve *niqqud*-izing existing datasets... with Machine Learning (!)
- Pairs of English suffixes are more or less frequent on a given base... **what if the difference of variability observed for e.g. *-ity* and *-ness* was due to different amounts of noise coming from frequency contrasts?** Appendix II shows some posthoc stats that tend to exclude this eventuality.

Caveats and avenue for future work

- **Written Hebrew, being usually devoid of vowels, is characterized by a high degree of ambiguity!**
- We tried to control for this by using maximally unambiguous forms (e.g. plural). Two potential alternatives:
  - **Use contextual word embeddings to disambiguate**. However, this relocates the issue in the choice of a "suitable" context for each target word.
  - **Train models on textual data including vowels markings** (*niqqud*). This would probably involve *niqqud*-izing existing datasets... with Machine Learning (!)
- Pairs of English suffixes are more or less frequent on a given base... **what if the difference of variability observed for e.g. *-ity* and *-ness* was due to different amounts of noise coming from frequency contrasts?** Appendix II shows some posthoc stats that tend to exclude this eventuality.

Caveats and avenue for future work

- **Written Hebrew, being usually devoid of vowels, is characterized by a high degree of ambiguity!**
- We tried to control for this by using maximally unambiguous forms (e.g. plural). Two potential alternatives:
    - **Use contextual word embeddings to disambiguate**. However, this relocates the issue in the choice of a "suitable" context for each target word.
    - **Train models on textual data including vowels markings** (*niqqud*). This would probably involve *niqqud*-izing existing datasets... with Machine Learning (!)
- Pairs of English suffixes are more or less frequent on a given base... **what if the difference of variability observed for e.g. *-ity* and *-ness* was due to different amounts of noise coming from frequency contrasts?** Appendix II shows some posthoc stats that tend to exclude this eventuality.

Thank you!

And special thanks to: Roger Levy, Adam Albright, Michael Elhadad.

Selected references I

[1] Maya Arad. "Locality Constraints on the Interpretation of Roots: The Case of Hebrew Denominal Verbs". In: *Natural Language and Linguistic Theory* 21.4 (Nov. 2003), pp. 737–778. DOI: 10.1023/a:1025533719905. URL: https://doi.org/10.1023/a:1025533719905.

[2] Piotr Bojanowski et al. "Enriching Word Vectors with Subword Information". In: *arXiv preprint arXiv:1607.04606* (2016). DOI: 10.48550/arXiv.1607.04606. URL: https://arxiv.org/abs/1607.04606.

[3] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). DOI: 10.48550/arXiv.1810.04805. arXiv: 1810.04805.

[4] Edouard Grave et al. "Learning Word Vectors for 157 Languages". In: *CoRR* abs/1802.06893 (2018). DOI: 10.48550/arXiv.1802.06893. arXiv: 1802.06893.

[5] Louis Guttman. "Some necessary conditions for common-factor analysis". In: *Psychometrika* 19.2 (June 1954), pp. 149–161. DOI: 10.1007/bf02289162. URL: https://doi.org/10.1007/bf02289162.

[6] Morris Halle and Alec Marantz. "Distributed Morphology and the Pieces of Inflection". In: *The View from Building 20*. Cambridge, MA: MIT Press, 1993, pp. 111–176.

Selected references II

[7] D. Jurafsky and J.H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in artificial intelligence. Pearson Prentice Hall, 2000. ISBN: 9780131873216.

[8] Paul Kiparsky. "Remarks on denominal verbs". In: *Argument Structure*. Ed. by A. Alsina, J. Bresnan, and P. Sells. Stanford: CLSI, 1997, pp. 473–499.

[9] Paul Kiparsky. "Word Formation and the Lexicon". In: *Proceedings of the Mid-America Linguistics Conference, University of Kansas*. Ed. by Fred Ingeman. 1982, pp. 3–29.

[10] Alec Marantz. "Roots: the universality of root and pattern morphology". In: *conference on Afro-Asiatic languages, University of Paris VII*. Vol. 3. 2000, p. 14.

[11] Tomas Mikolov et al. "Advances in Pre-Training Distributed Word Representations". In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

Selected references III

[12] Tomás Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2013. DOI: 10.48550/arXiv.1301.3781. URL: http://arxiv.org/abs/1301.3781.

[13] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

[14] Amit Seker et al. "AlephBERT: A Hebrew Large Pre-Trained Language Model to Start-off your Hebrew NLP Application With". In: *CoRR* abs/2104.04052 (2021). DOI: 10.48550/arXiv.2104.04052. arXiv: 2104.04052.

Appendix I: Hebrew

Data generation procedure

- Elaborate a list of nominal templates with templatic consonants, and match those templates against nouns extracted from the PoS-tagged Knesset Meetings Corpus, to **obtain a list of nouns with templatic consonants**.
- For each noun N of this list:
  - Extract its root (easy because we know its template!), and **generate candidate root-derived verbs** $(V_i)_{i\in[1,K]}$ using the verbal templates from Table 3 (next slide).
  - From the noun itself, **generate candidate denominal verbs**[2]using the template mapping in Table 5 (next slide).
- Match the candidate forms (and any inflected variant thereof) against the corpus to **filter unattested elements**.
- Manually inspect the remaining candidates.

[2] Note that one given noun can in practice give rise to several denominal forms, because certain nominal templates are compatible with more than one denominal template, see e.g. row 2 of Table 5.

Appendix I: Hebrew

General testing strategy for Hebrew data

- **Generate** a dataset of n $(N, (V_i)_{i \in [1,K]}, D)$ triplets.
- **Embed** and **reduce** the dimensionality of the data to get vectors that are as meaningful and noiseless as possible.
- **Compute** $CosSim(\overrightarrow{N}, \overrightarrow{D})$ and $\max_i CosSim(\overrightarrow{N}, \overrightarrow{V_i})$ / $\frac{1}{K} \sum_{i=1}^{K} CosSim(\overrightarrow{N}, \overrightarrow{V_i})$, for each triplet, to get a list of n pairs of scores.
- **Perform** a one-tailed Wilcoxon test for matched-pairs on the data and compute the relevant effect sizes. We used Cliff's Δ because it is a robust, non-parametric measure that ended up being a bit more stringent than Cohen's d in our case.

Appendix I: Hebrew

Verbal templates
CaCaC
niCCaC
CiCCeC
CuCCaC
hiCCiC
huCCaC
hitCaCCeC

Table 3: Verbal templates susceptible to apply at the root level

Step	# datapoints
Generation from templates	**1435**
Filtering via corpus	1435-1322 = **113**
Manual inspection	113-47 = **66**

Table 4: Number of datapoints at each step of the generation procedure

Nominal template	Denominal template(s)
tiCCoCet tiCCoCa taCCiC	letaCCeC
CeCCon	leCaCCen lehitCaCCen
maCCeC miCCeCet miCCaC	lemaCCeC lehitmaCCeC
šaCCeCet	lešaCCeC lehištaCCeC
CaCaCat	leCaCCet lehitCaCCet

Table 5: Correspondence between nominal templates involving templatic consonants and the denominal (verbal) template that can apply on top of them

Appendix I: Hebrew

Construction/collection of the word embedding models

- 4 architectures: Word2Vec [12], GloVe [13], fastText [2], BERT [3]:
  - fastText [4] and BERT (AlephBERT, [14]) were pretrained.
  - Word2Vec and GloVe were trained on Hebrew Wikipedia dumps. GloVe was trained in 2 dimensions: 50 and 100.
- Dimension reduction was performed on the data using PCA along with the Guttman-Kaiser criterion [5] to determine the optimal reduced dimension.

Model	Word2Vec	GloVe	fastText	BERT
# vectors	584 160	584 162	2 billion	NA
Initial dimension	100	50/100	300	768
PCA-reduced dimension	27	28/46	50	107

Table 6: Characteristics of the models

Appendix II: English

Data generation procedure

- Merge two Python lexicons: **NLTK** (236736 words) and **english-words** (25487 words), for a total of 240788 words.
- Given two suffixes s_1 and s_2:
  - find the words ending with s_1 in the lexicon;
  - replace s_1 by s_2;
  - if the newly formed word is also present in the lexicon (modulo a few character changes), **add the triplet ($\boldsymbol{b}$, $\boldsymbol{b}$-s_1, $\boldsymbol{b}$-s_2) to the dataset**.
- This generated 683 triplets for *-ity*/*-ness* and 555 triplets for textit-al/*-less*, that we manually filtered.
- Triplets for which at least one element was not “embeddable” were also automatically excluded.

Appendix II: English

Characteristics of the pretrained models

- We chose static embedding with matching initial dimensions. BERT's initial dimension could not be lower tan 768.
- Dimension was reduced by fitting PCA on the relevant datasets, and retaining 90% of the explained variance.

Model		Word2Vec	GloVe	fastText	BERT
Pretrained on		Google News (100B words)	Common Crawl (840B tokens)	Common Crawl (600B tokens)	BookCorpus +Wikipedia (2.5+0.8B words)
Initial dimension		300	300	300	768
PCA-reduced dimension	-ity/-ness	52	129	130	198
	-al/-less	32	79	84	152

Appendix II: English

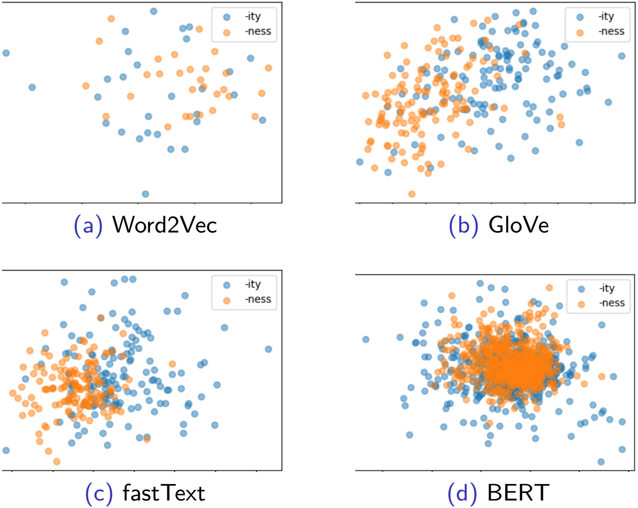


Figure 5: 2D PCA of the $\overrightarrow{\textit{-ity}}$ and $\overrightarrow{\textit{-ness}}$ vectors

Appendix II: English

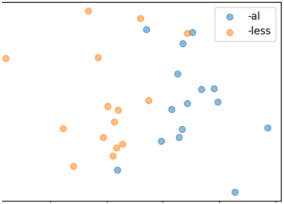
Results for English *-al*/*-less* suffixes

- Results overall less significant than for the *-ity*/*-ness* pair.
- For Word2Vec however, the size of the dataset (15) is too small, which questions the relevance of the negative result for this model.

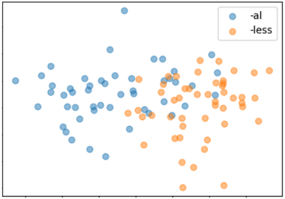
	Word2Vec	GloVe	fastText	BERT
n	15	49	54	205
"dispersion"	.054 .12 (N)	7e-108 .53 (L)	1e-11 .11 (N)	0 .47 (M)
"variation"	.49 .29 (S)	1e-9 .65 (L)	.17 .16 (S)	9e-26 .65 (L)

Table 7: *p*-values (2-tailed Wilcoxon) and effect sizes (Cliff's Δ; N=Negligible; S=Small; M=Medium; L=Large) for dispersion and variation measures and 4 embedding models

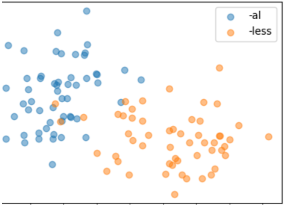
Appendix II: English



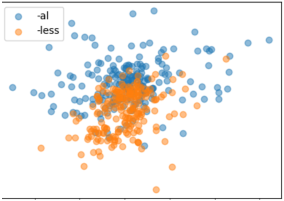
(a) Word2Vec



(b) GloVe



(c) fastText



(d) BERT

Figure 6: 2D PCA of the $\overrightarrow{\textit{-al}}$ and $\overrightarrow{\textit{-less}}$ vectors

Appendix II: English

The frequency confound (thanks to Adam Albright!)

- A potential confound in the comparison of two suffixes s_1 and s_2 (e.g. *-ity* and *-ness*) might be a difference in frequency between *a*-s_1 and *a*-s_2 for a given adjective *a*.
- Indeed, less occurrences of a given word may lead a neural model to derive a noisier representation, independently of linguistic theory.
- This would be a big problem if *-ity* and *-al* (predicted to be more variable *in theory*), also happened to be less frequent (and hence, potentially noisier).

Appendix II: English

Posthoc frequency analysis

- The Table below gathers statistics about the frequency ratios between *a-ity* and *a-ness* (frequencies extracted from Wikipedia by IlyaSemenov on GitHub).
- *-ity* is more frequent than *-ness* on a given base 4 to 5 times *more often*; and when it is the case the discrepancy in frequency is also more drastic!
- Suggests that the frequency contrast in the case of *-ity* and *-ness* does not go in the “confounding” direction!

		Word2Vec	GloVe	fastText	BERT
f-ratios favoring ity	n	23	93	103	122
	mean	298	1379	1280	1278
	median	68	100	100	120
f-ratios favoring ness	n	6	19	21	32
	mean	41	108	30	72
	median	3	5	5	5
not computed		0	14	20	456