# Word-Embeddings Distinguish Denominal and Root-Derived Verbs in Semitic

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

[1]Massachusetts Institute of Technology

August 16, 2022

## Full disclaimer

- Thank you so much for having us!
- I (Adèle) am here to present this work. I am not a native speaker of Modern Hebrew, but my two co-authors, Omri and Ido, are. I will do my best to answer Hebrew-related questions!
- This talk will focus on the bridges between generative linguistics and machine learning. Not a lot of logical background...sorry in advance!
- We would like to thank Roger Levy from MIT Brain and Cognitive Sciences, who helped us develop this project as part of the Computational Psycholinguistics class.

# Introduction: Hebrew morphology and the two-level model

# A few basic principles of word formation

## Morphology and semantic/phonological transparency

- Some but not all compounds have a compositional meaning: *(huckle$_{?}$-berry)* vs *black$_{\blacksquare}$-berry*/*blue$_{\blacksquare}$-berry* [1].
- Some but not all English suffixes leave stress intact: *glóbal* $\rightarrow$ *glóbal-ness*, but *globál-ity*.

# A few basic principles of word formation

## Morphology and semantic/phonological transparency

- Some but not all compounds have a compositional meaning: *(huckle?-berry)* vs *black■-berry/blue■-berry* [1].
- Some but not all English suffixes leave stress intact: *glóbal* → *glóbal-ness*, but *globál-ity*.

## The two-level model ([2], [3] a.o.)

- Morphological operations can be of two types...
  - **Level 1**: idiosyncratic, non-compositional, below-word.
  - **Level 2**: deterministic, compositional, above-word.

# A few basic principles of word formation

## Morphology and semantic/phonological transparency

- Some but not all compounds have a compositional meaning: *(huckle*$_{?}$*-berry)* vs *black*$_{\blacksquare}$*-berry*/*blue*$_{\blacksquare}$*-berry* [1].
- Some but not all English suffixes leave stress intact: *glóbal* $\rightarrow$ *glóbal-ness*, but *globál-ity*.

## The two-level model ([2], [3] a.o.)

- Morphological operations can be of two types...
  - **Level 1**: idiosyncratic, non-compositional, below-word.
  - **Level 2**: deterministic, compositional, above-word.
- A *word* is created once a root ($\sqrt{}$) is merged with a functional head: *n*(ominalizer), *v*(erbalizer), *a*(djectivizer) etc.
- The first head to be merged sets the rough semantic/phonological features of the newly created word.

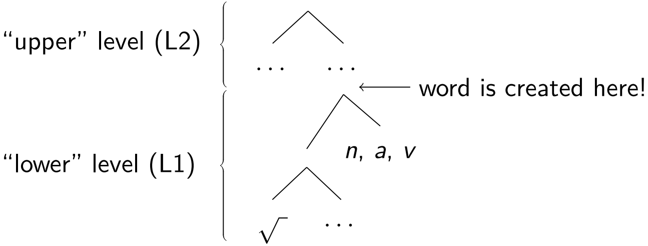



Figure 1: Two-level morphology

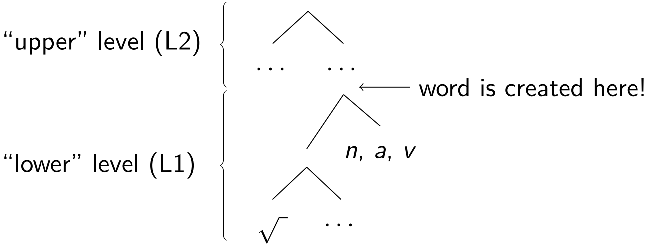



Figure 1: Two-level morphology

## Key semantic predictions of the two-level model

We focus on the semantic effects of word-formation (L1) and subsequent affixation (L2). Two key predictions:

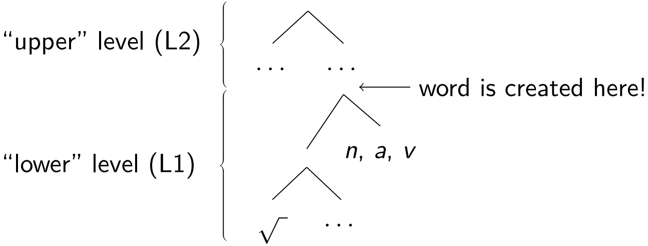



Figure 1: Two-level morphology

## Key semantic predictions of the two-level model

We focus on the semantic effects of word-formation (L1) and subsequent affixation (L2). Two key predictions:

- A **Words derived from the same *root* via L1 operations may arbitrarily differ semantically.**

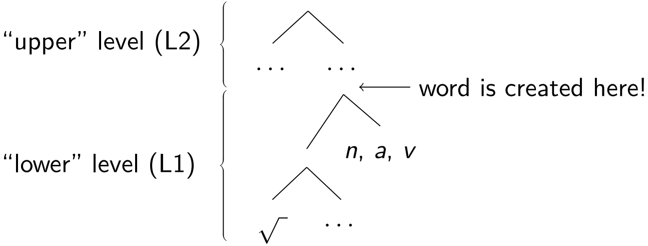



Figure 1: Two-level morphology

## Key semantic predictions of the two-level model

We focus on the semantic effects of word-formation (L1) and subsequent affixation (L2). Two key predictions:

- A **Words derived from the same *root* via L1 operations may arbitrarily differ semantically.**
- B **Words derived from the same *base word via* L2 operations should be closely related semantically.**

# Application to Semitic (templatic) morphology

## A non-concatenative system

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

# Application to Semitic (templatic) morphology

## A non-concatenative system

- **In Modern Hebrew (MH), 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

- 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'.

# Application to Semitic (templatic) morphology

## A non-concatenative system

- **In Modern Hebrew (MH), 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

- 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*.

# Application to Semitic (templatic) morphology

## A non-concatenative system

- **In Modern Hebrew (MH), 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

- 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**.

# Case study: Hebrew denominal verbs

# 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 (L1), followed by that of a *v*-head (L2).

# 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 (L1), followed by that of a *v*-head (L2).
- It is not easy to tease apart denominals from “basic” verbs derived directly from a root in English corpora (but see [4]).

# 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 (L1), followed by that of a *v*-head (L2).
- It is not easy to tease apart denominals from "basic" verbs derived directly from a root in English corpora (but see [4]).
- **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...

# 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 (L1), followed by that of a *v*-head (L2).
- It is not easy to tease apart denominals from "basic" verbs derived directly from a root in English corpora (but see [4]).
- **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...

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

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

## Denominal *vs* root-derived verbs [5]

- Back to the predictions of the 2-level model...

## Denominal *vs* root-derived verbs [5]

- 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 L1 operation), we expect them to differ semantically in a somewhat arbitrary way.

## Denominal *vs* root-derived verbs [5]

- 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 L1 operation), we expect them to differ semantically in a somewhat arbitrary way.
  - (B) If a denominal $D$ derives from a base noun $N$ (*via* a L2 operation), we expect them to be close semantically.

## Denominal *vs* root-derived verbs [5]

- 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 L1 operation), we expect them to differ semantically in a somewhat arbitrary way.
  - B. If a denominal $D$ derives from a base noun $N$ (*via* a L2 operation), we expect them to be close semantically.
- Thus, given a root $\sqrt{}$, a noun $N$, a verb $V$, a denominal $D$, s.t. $\sqrt{} \stackrel{L1}{\rightarrow} N$, $\sqrt{} \stackrel{L1}{\rightarrow} V$, and $N \stackrel{L2}{\rightarrow} D$, we expect:

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

For some well-chosen semantic measure $\mathcal{S}$ between pairs of words.

## Denominal *vs* root-derived verbs [5]

- 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 L1 operation), we expect them to differ semantically in a somewhat arbitrary way.
  - B. If a denominal $D$ derives from a base noun $N$ (*via* a L2 operation), we expect them to be close semantically.
- Thus, given a root $\sqrt{}$, a noun $N$, a verb $V$, a denominal $D$, s.t. $\sqrt{} \xrightarrow{L1} N$, $\sqrt{} \xrightarrow{L1} V$, and $N \xrightarrow{L2} D$, 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:

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

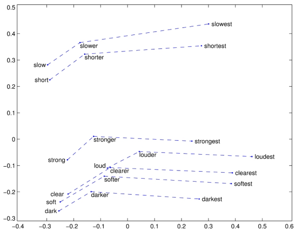
# Modeling the predictions within Hebrew word embedding models

## Relevance of word embeddings to our task

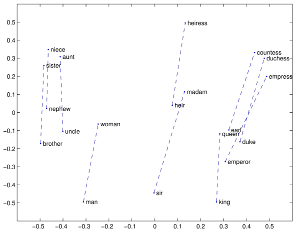
- Word embeddings are high-dimensional vector representations of words, often learned as "byproducts" of ML-related tasks (word prediction, classification...) [6].

## Relevance of word embeddings to our task

- Word embeddings are high-dimensional vector representations of words, often learned as "byproducts" of ML-related tasks (word prediction, classification...) [6].
- **Past empirical evidence in favor of embeddings' encoding of semantic features and relationships** [7].
- Embeddings come with a robust measure of semantic similarity, cosine similarity!




(a) Positive form → comparative → superlative transformations [7]




(b) Masculine ↔ feminine transformations [7]

## 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:

[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 L1 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. $\sqrt{\ } \xrightarrow{L1} A$, and $\sqrt{\ } \xrightarrow{L1} B$, we expect $\overrightarrow{A}$ and $\overrightarrow{B}$ to be randomly distributed across $Area(\sqrt{\ })$.

[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 L1 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. $\sqrt{} \xrightarrow{L1} A$, and $\sqrt{} \xrightarrow{L1} B$, we expect $\overrightarrow{A}$ and $\overrightarrow{B}$ to be randomly distributed across $Area(\sqrt{})$.
  - (B) Given $\sqrt{}$, $A$ and $B$, s.t. $\sqrt{} \xrightarrow{L1} A \xrightarrow{L2} B$, we expect $\overrightarrow{A}$ and $\overrightarrow{B}$ to be very close to each other within $Area(\sqrt{})$.

---

[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 L1 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 $\{\vec{X} | \surd \to^* X\}$. The predictions of the 2-level model become:
  - (A) Given a root $\surd$, and $A$, $B$, s.t. $\surd \overset{L1}{\to} A$, and $\surd \overset{L1}{\to} B$, we expect $\vec{A}$ and $\vec{B}$ to be randomly distributed across $Area(\surd)$.
  - (B) Given $\surd$, $A$ and $B$, s.t. $\surd \overset{L1}{\to} A \overset{L2}{\to} B$, we expect $\vec{A}$ and $\vec{B}$ to be very close to each other within $Area(\surd)$.
- Let $\surd$, $N$, $D$, $(V_i)_{i \in [1,K]}$, be s.t. $\surd \overset{L1}{\to} N$, $\forall i \in [1, K]$ $\surd \overset{L1}{\to} V_i$, and $N \overset{L2}{\to} D$. We predict:

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

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

[1] The stronger hypothesis is not expected to hold all the time, because the closest $\vec{V_i}$ may accidentally end up closer to $\vec{N}$ than $\vec{D}$ is, due to the arbitrariness of L1 operations. This motivates the use of the weaker hypothesis.

# Testing the predictions within Hebrew word embedding models

## Testing strategy

- **Generate** a dataset of $n$ $(N, (V_i)_{i \in [1,K]}, D)$ triplets.

## Testing strategy

- **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.

## Testing strategy

- **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.

## Testing strategy

- **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.

## 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**.

[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 3.

## 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 1 (next slide).

[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 3.

## 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 1 (next slide).
  - From the noun itself, **generate candidate denominal verbs**[2] using the template mapping in Table 3 (next slide).

[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 3.

## 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 1 (next slide).
  - From the noun itself, **generate candidate denominal verbs**[2] using the template mapping in Table 3 (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 3.

| Verbal templates |
|---|
| CaCaC |
| niCCaC |
| CiCCeC |
| CuCCaC |
| hiCCiC |
| huCCaC |
| hitCaCCeC |

Table 1: Verbal templates susceptible to apply at the root level

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

Table 2: Number of data points at each step of the generation procedure

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

Table 3: Correspondence between nominal templates involving templatic consonants and the denominal (verbal) template that can apply on top of them

## Preparation of the word embeddings

- 4 models: Word2Vec [8], GloVe [7], fastText [9], BERT [10]:
  - fastText [11] and BERT (AlephBERT, [12]) were pretrained.[3]
  - Word2Vec and GloVe were trained on Hebrew Wikipedia dumps. GloVe was trained with two initial dimensions: 50 and 100.

[3] To get embeddings in the BERT model, we chose to sum the last 4 layers obtained after a forward pass performed on a single tokenized input (word). No context was provided.

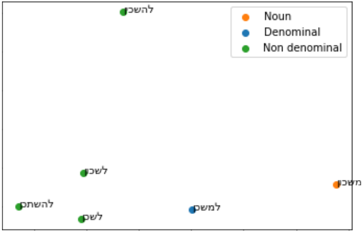
## Preparation of the word embeddings

- 4 models: Word2Vec [8], GloVe [7], fastText [9], BERT [10]:
  - fastText [11] and BERT (AlephBERT, [12]) were pretrained.[3]
  - Word2Vec and GloVe were trained on Hebrew Wikipedia dumps. GloVe was trained with two initial dimensions: 50 and 100.
- Dimension reduction was performed on the data using PCA along with the Guttman-Kaiser criterion [13] to determine the optimal reduced dimension.

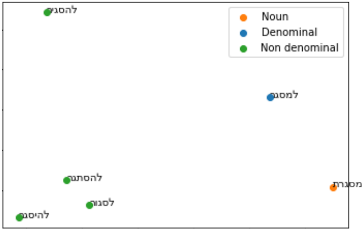
| Model | Word2Vec | GloVe | fastText | BERT |
|---|---|---|---|---|
| # vectors | 584 160 | 584 162 | 2 billion | NA |
| Initial dimension | 100 | 50/100 | 300 | 768 |
| PCA-reduced dimension | 27 | 28/46 | 50 | 107 |

Table 4: Characteristics of the models

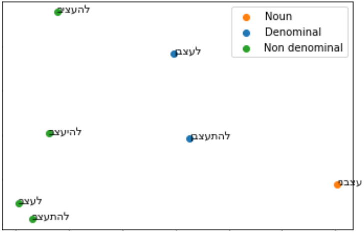
[3] To get embeddings in the BERT model, we chose to sum the last 4 layers obtained after a forward pass performed on a single tokenized input (word). No context was provided.



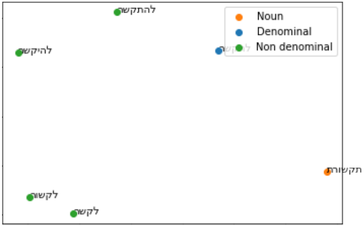

(a) Noun: ‘pawning’; Denominal: ‘to pawn’




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




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




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

Figure 3: 2D-reduction of a few data points (PCA, cosine kernel, fastText)

## Results

- **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 ($\text{GloVe}_{50}$, BERT) are significant.
  - Large effect sizes on the significant results, except on $\text{GloVe}_{100}$.

| | Word2Vec | $\text{GloVe}_{50}$ | $\text{GloVe}_{100}$ | fastText | AlephBERT |
|---|---|---|---|---|---|
| # data points | 31 | 31 | 31 | 53 | 66 |
| Weak hyp. (mean) | $1.06 \times 10^{-6}$ 0.86 (Large) | $2.43 \times 10^{-4}$ 0.52 (Large) | $6.64 \times 10^{-5}$ 0.66 (Large) | $1.42 \times 10^{-10}$ 0.79 (Large) | $4.84 \times 10^{-4}$ 0.30 (Small) |
| Strong hyp. (max) | $3.77 \times 10^{-5}$ 0.66 (Large) | $1.68 \times 10^{-1}$ 0.06 (Negligible) | $2.87 \times 10^{-2}$ 0.20 (Small) | $1.39 \times 10^{-8}$ 0.62 (Large) | $3.59 \times 10^{-1}$ 0.02 (Negligible) |

Table 5: $p$-values and effect sizes (Cliff's $\Delta$) for the weak and strong hypotheses and 4 embedding models

## Conclusion

- Weak hypothesis verified on all models, **robust prediction**!

## Conclusion

- Weak hypothesis verified on all models, **robust prediction**!
- What is going on with $\text{GloVe}_{50}$ and BERT and the stronger hypothesis?

## Conclusion

- Weak hypothesis verified on all models, **robust prediction**!
- What is going on with $\text{GloVe}_{50}$ and BERT and the stronger hypothesis?
  - First, recall that the stronger hypothesis was "noisier" because it could be accidentally violated for some triplets, due to the arbitrariness of L1 operations.

## Conclusion

- Weak hypothesis verified on all models, **robust prediction**!
- What is going on with $\text{GloVe}_{50}$ and BERT and the stronger hypothesis?
  - First, recall that the stronger hypothesis was “noisier” because it could be accidentally violated for some triplets, due to the arbitrariness of L1 operations.
  - $\text{GloVe}_{50}$ may have been too impoverished from the beginning (**low dimensionality during training**)... this explains why $\text{GloVe}_{100}$ manages to reach significance.

## Conclusion

- Weak hypothesis verified on all models, **robust prediction**!
- What is going on with $\text{GloVe}_{50}$ and BERT and the stronger hypothesis?
  - First, recall that the stronger hypothesis was "noisier" because it could be accidentally violated for some triplets, due to the arbitrariness of L1 operations.
  - $\text{GloVe}_{50}$ may have been too impoverished from the beginning (**low dimensionality during training**)... this explains why $\text{GloVe}_{100}$ manages to reach significance.
  - But then, how about BERT, which had the highest initial dimensionality? **BERT may have performed poorly because it was not used at its full potential** (i.e. with context words)!

## Caveats, future work, new issues

- **Written Hebrew, being usually devoid of vowels, is characterized by a high degree of ambiguity!**

## Caveats, future work, new issues

- **Written Hebrew, being usually devoid of vowels, is characterized by a high degree of ambiguity!**
- This certainly adds significant noise to the word vectors produced by static embeddings – even though it is unclear how this noise influences our hypotheses.

## Caveats, future work, new issues

- **Written Hebrew, being usually devoid of vowels, is characterized by a high degree of ambiguity!**
- This certainly adds significant noise to the word vectors produced by static embeddings – even though it is unclear how this noise influences our hypotheses.
- We tried to control for this by using maximally unambiguous forms (e.g. by adding plural inflections). There are however two obvious alternatives to this "trick":

## Caveats, future work, new issues

- **Written Hebrew, being usually devoid of vowels, is characterized by a high degree of ambiguity!**
- This certainly adds significant noise to the word vectors produced by static embeddings – even though it is unclear how this noise influences our hypotheses.
- We tried to control for this by using maximally unambiguous forms (e.g. by adding plural inflections). There are however two obvious alternatives to this "trick":
  - **Use contextual word embeddings properly**. But this relocates the issue in the choice of a "suitable" context for each target word (subjective task!). This moreover requires to deal with varying (uncontrolled!) argument structures.

## Caveats, future work, new issues

- **Written Hebrew, being usually devoid of vowels, is characterized by a high degree of ambiguity!**
- This certainly adds significant noise to the word vectors produced by static embeddings – even though it is unclear how this noise influences our hypotheses.
- We tried to control for this by using maximally unambiguous forms (e.g. by adding plural inflections). There are however two obvious alternatives to this "trick":
  - **Use contextual word embeddings properly**. But this relocates the issue in the choice of a "suitable" context for each target word (subjective task!). This moreover requires to deal with varying (uncontrolled!) argument structures.
  - **Train models on textual data including vowels markings** (called *niqqud*). This would probably involve *niqqud*-izing existing datasets… with ML! Again, this solution only moves the problem (disambiguation) elsewhere in the pipeline.

# Thank you!

# Selected references I


- M. Aronoff, *Word Formation in Generative Grammar*, ser. Linguistic Inquiry monographs. MIT press, 1976, ISBN: 9780262510172.
- M. Halle and A. Marantz, "Distributed morphology and the pieces of inflection," in *The View from Building 20*, Cambridge, MA: MIT Press, 1993, pp. 111–176.
- A. Marantz, "Roots: The universality of root and pattern morphology," in *conference on Afro-Asiatic languages, University of Paris VII*, vol. 3, 2000, p. 14.
- P. Kiparsky, "Remarks on denominal verbs," in *Argument Structure*, A. Alsina, J. Bresnan, and P. Sells, Eds., Stanford: CLSI, 1997, pp. 473–499.
- M. Arad, "Locality constraints on the interpretation of roots: The case of hebrew denominal verbs," *Natural Language and Linguistic Theory*, vol. 21, no. 4, pp. 737–778, Nov. 2003. DOI: 10.1023/a:1025533719905. [Online]. Available: https://doi.org/10.1023/a:1025533719905.
- D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, ser. Prentice Hall series in artificial intelligence. Pearson Prentice Hall, 2000, ISBN: 9780131873216.

# Selected references II

J. Pennington, R. Socher, and C. 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. [Online]. Available: https://aclanthology.org/D14-1162.

T. Mikolov, K. Chen, G. Corrado, and J. Dean, "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*, Y. Bengio and Y. LeCun, Eds., 2013. DOI: 10.48550/arXiv.1301.3781. [Online]. Available: http://arxiv.org/abs/1301.3781.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016. DOI: 10.48550/arXiv.1607.04606. [Online]. Available: https://arxiv.org/abs/1607.04606.

# Selected references III


J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. DOI: 10.48550/arXiv.1810.04805. arXiv: 1810.04805.

E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," *CoRR*, vol. abs/1802.06893, 2018. DOI: 10.48550/arXiv.1802.06893. arXiv: 1802.06893.

A. Seker, E. Bandel, D. Bareket, I. Brusilovsky, R. S. Greenfeld, and R. Tsarfaty, "Alephbert: A hebrew large pre-trained language model to start-off your hebrew NLP application with," *CoRR*, vol. abs/2104.04052, 2021. DOI: 10.48550/arXiv.2104.04052. arXiv: 2104.04052.

L. Guttman, "Some necessary conditions for common-factor analysis," *Psychometrika*, vol. 19, no. 2, pp. 149–161, Jun. 1954. DOI: 10.1007/bf02289162. [Online]. Available: https://doi.org/10.1007/bf02289162.