The logical form of lexical semantics

Itamar Kastner

ESSLLI 2022, Session 4

Last time

- Another empirical domain: Levinson's root classes
- Analysis: root types
- Analysis: functional primitives

Today

- Experimental approaches
- Computational approaches
- Including some of my work in progress

What are we modelling?

"Quantitative", "empirical" and "experimental" are very broad terms.
We might ask:

- Are the empirical patterns robust in a given language?
- Are the empirical patterns robust across languages?
- Do language users generalize them to novel items?
- Do the linguistic patterns have behavioural correlates?
- Can they be predicted from distributional patterns?

We'll do a bit of each.

1. Recap
2. Experimental: Are the patterns robust?
   - Levinson
   - Changes of state
3. Experimental: Do they generalize?
4. Computational: Can they be derived from distributions?
5. Summary

Levinson

Irwin and Kastner (2020) wanted to test Levinson's generalizations in a large-scale study.
We added a control condition and two new diagnostics to Levinson's three tests.

Six conditions

Contexts:

1. Transitive (baseline/control)
2. Pseudo-Res (Levinson 2007, 2010, 2014)
3. Double object (Clark and Clark 1979; Levinson 2014)
4. No Theme (Levin and Rappaport Hovav 2005; Levinson 2007, 2014)
5. Participle + *still* (Kratzer 2000; Anagnostopoulou 2015)
6. Gerund (Roßdeutscher and Kamp 2010; Alexiadou et al. 2017)

64 participants online.

Levinson

	Transitive	PseudoRes	Double Obj	No Theme	Participle + *still*	Gerund
Root *pile*	✓ Ray piled the cush-ions	✓ Ray piled the cush-ions high	✗ Ray piled his mom some cush-ions	✗ Ray was piling all day	✓ Ray piled some cushions yesterday, and they were still piled today	✓ The high piling of the cush-ions was a success
Exp *bake*	✓ Ray baked some cook-ies	✗ Ray baked the ingre-dients tasty	✓ Ray baked his mom some cookies	✓ Ray was bak-ing all day yes-terday	✗ Ray baked some cookies yesterday, and they were still baked today	✗ The tasty baking of the ingre-dients was a success
COS *open*	✓ Ray opened the door	✗ Ray opened the door tiny	✓ Ray opened his buddy a beer	✗ Ray was open-ing all day	✓ Ray opened the window yesterday, and it was still opened	✗ The wide opening of the window both-ered

Levinson

	Transitive	PseudoRes	Double Obj	No Theme	Participle + *still*	Gerund
Root *pile*	✓ Ray piled the cush-ions	✓ Ray piled the cush-ions high	✗ Ray piled his mom some cush-ions	✗ Ray was piling all day	✓ Ray piled some cushions yesterday, and they were still piled today	✓ The high piling of the cush-ions was a success
Exp *bake*	✓ Ray baked some cook-ies	✗ Ray baked the ingre-dients tasty	✓ Ray baked his mom some cookies	✓ Ray was bak-ing all day yes-terday	✗ Ray baked some cookies yesterday, and they were still baked today	✗ The tasty baking of the ingre-dients was a success
COS *open*	✓ Ray opened the door	✗ Ray opened the door tiny	✓ Ray opened his buddy a beer	✗ Ray was open-ing all day	✓ Ray opened the window yesterday, and it was still opened	✗ The wide opening of the window both-ered

Levinson

	Transitive	PseudoRes	Double Obj	No Theme	Participle + *still*	Gerund
Root *pile*	✓ Ray piled the cush-ions	✓ Ray piled the cush-ions high	✗ Ray piled his mom some cush-ions	✗ Ray was piling all day	✓ Ray piled some cushions yesterday, and they were still piled today	✓ The high piling of the cush-ions was a success
Exp *bake*	✓ Ray baked some cook-ies	✗ Ray baked the ingre-dients tasty	✓ Ray baked his mom some cookies	✓ Ray was bak-ing all day yes-terday	✗ Ray baked some cookies yesterday, and they were still baked today	✗ The tasty baking of the ingre-dients was a success
COS *open*	✓ Ray opened the door	✗ Ray opened the door tiny	✓ Ray opened his buddy a beer	✗ Ray was open-ing all day	✓ Ray opened the window yesterday, and it was still opened	✗ The wide opening of the window both-ered

Levinson

	Transitive	PseudoRes	Double Obj	No Theme	Participle + *still*	Gerund
Root *pile*	✓ Ray piled the cush-ions	✓ Ray piled the cush-ions high	✗ Ray piled his mom some cush-ions	✗ Ray was piling all day	✓ Ray piled some cushions yesterday, and they were still piled today	✓ The high piling of the cush-ions was a success
Exp *bake*	✓ Ray baked some cook-ies	✗ Ray baked the ingre-dients tasty	✓ Ray baked his mom some cookies	✓ Ray was bak-ing all day yes-terday	✗ Ray baked some cookies yesterday, and they were still baked today	✗ The tasty baking of the ingre-dients was a success
COS *open*	✓ Ray opened the door	✗ Ray opened the door tiny	✓ Ray opened his buddy a beer	✗ Ray was open-ing all day	✓ Ray opened the window yesterday, and it was still opened	✗ The wide opening of the window both-ered

Levinson

	Transitive	PseudoRes	Double Obj	No Theme	Participle + *still*	Gerund
Root *pile*	✓ Ray piled the cush-ions	✓ Ray piled the cush-ions high	✗ Ray piled his mom some cush-ions	✗ Ray was piling all day	✓ Ray piled some cushions yesterday, and they were still piled today	✓ The high piling of the cush-ions was a success
Exp *bake*	✓ Ray baked some cook-ies	✗ Ray baked the ingre-dients tasty	✓ Ray baked his mom some cookies	✓ Ray was bak-ing all day yes-terday	✗ Ray baked some cookies yesterday, and they were still baked today	✗ The tasty baking of the ingre-dients was a success
COS *open*	✓ Ray opened the door	✗ Ray opened the door tiny	✓ Ray opened his buddy a beer	✗ Ray was open-ing all day	✓ Ray opened the window yesterday, and it was still opened	✗ The wide opening of the window both-ered

Levinson

	Transitive	PseudoRes	Double Obj	No Theme	Participle + *still*	Gerund
Root *pile*	✓ Ray piled the cush-ions	✓ Ray piled the cush-ions high	✗ Ray piled his mom some cush-ions	✗ Ray was piling all day	✓ Ray piled some cushions yesterday, and they were still piled today	✓ The high piling of the cush-ions was a success
Exp *bake*	✓ Ray baked some cook-ies	✗ Ray baked the ingre-dients tasty	✓ Ray baked his mom some cookies	✓ Ray was bak-ing all day yes-terday	✗ Ray baked some cookies yesterday, and they were still baked today	✗ The tasty baking of the ingre-dients was a success
COS *open*	✓ Ray opened the door	✗ Ray opened the door tiny	✓ Ray opened his buddy a beer	✗ Ray was open-ing all day	✓ Ray opened the window yesterday, and it was still opened	✗ The wide opening of the window both-ered

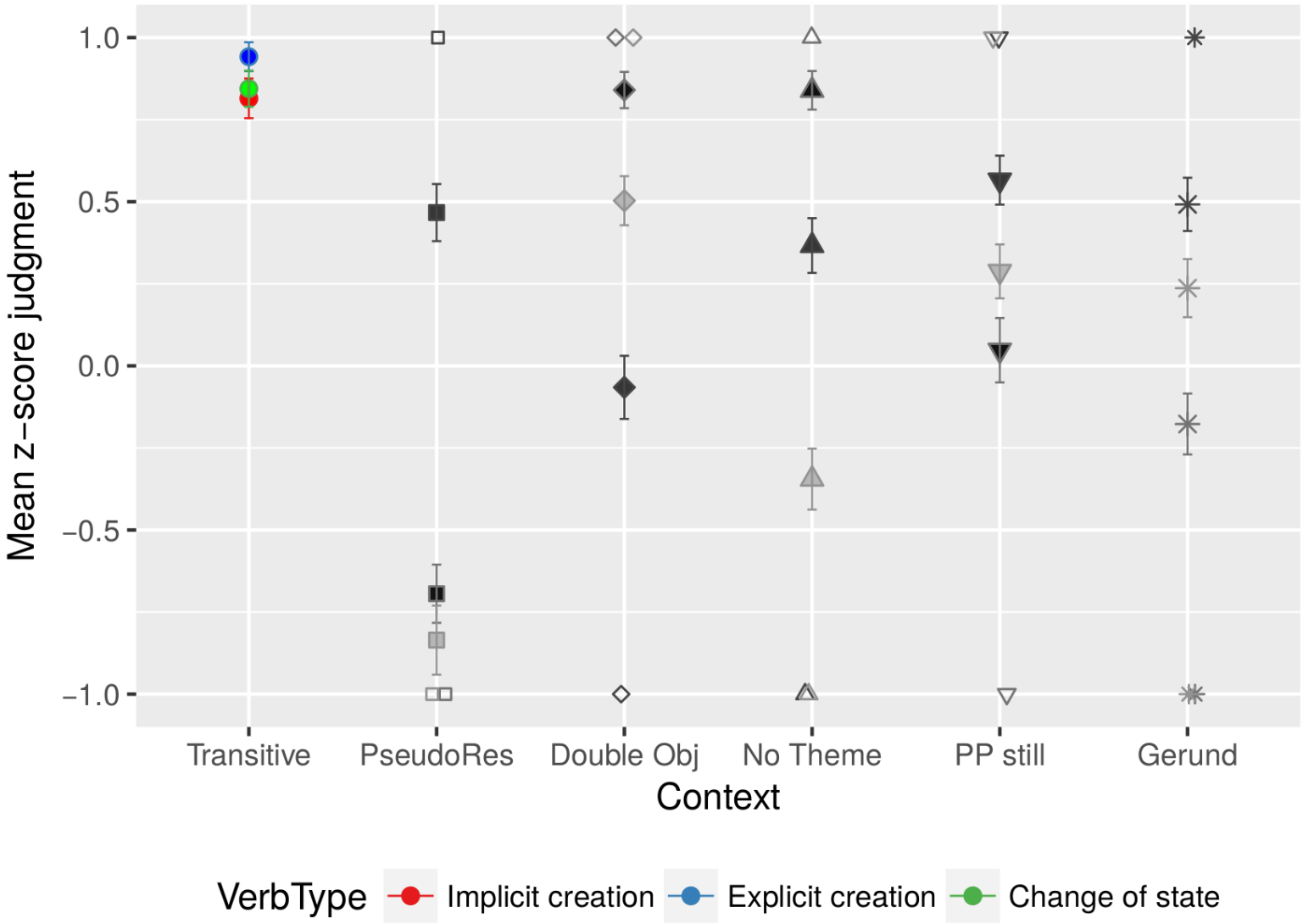
Levinson

1. Three verbs per Verb Type.

Root creation	Explicit creation	Change of state
pile	*bake*	*open*
stack	*build*	*clear*
braid	*cook*	*cool*

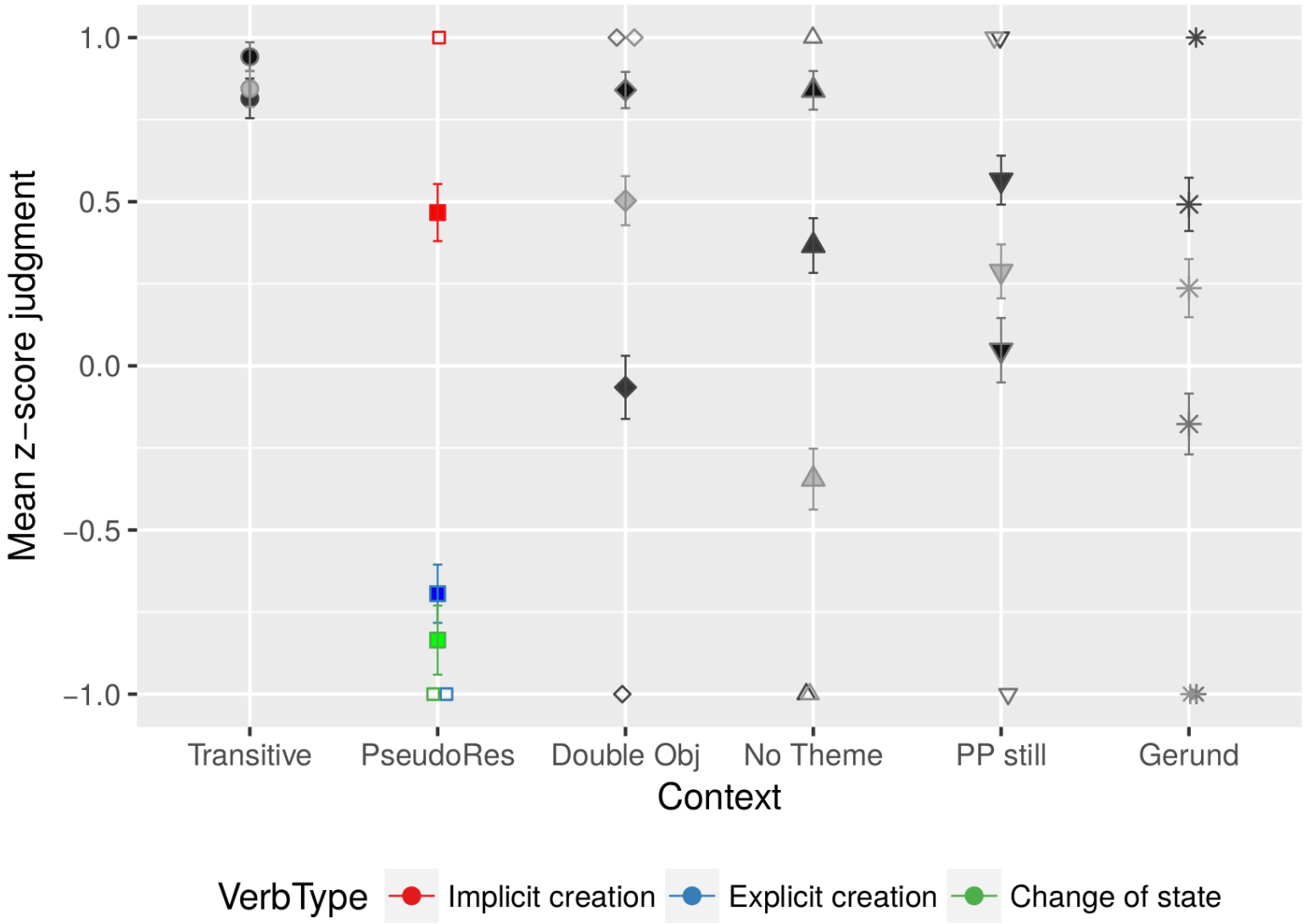
2. 3 verbs * 3 Types * 6 Contexts = 54 experimental items per subject.
3. Online judgment survey.
4. Ibex. (Drummond n.d)
5. Rating on a 7-point Likert scale, 1 “least acceptable” to 7 “most acceptable”.
6. Context responses were treatment coded with Transitive as reference.
7. Verb Type was contrast coded.

Levinson



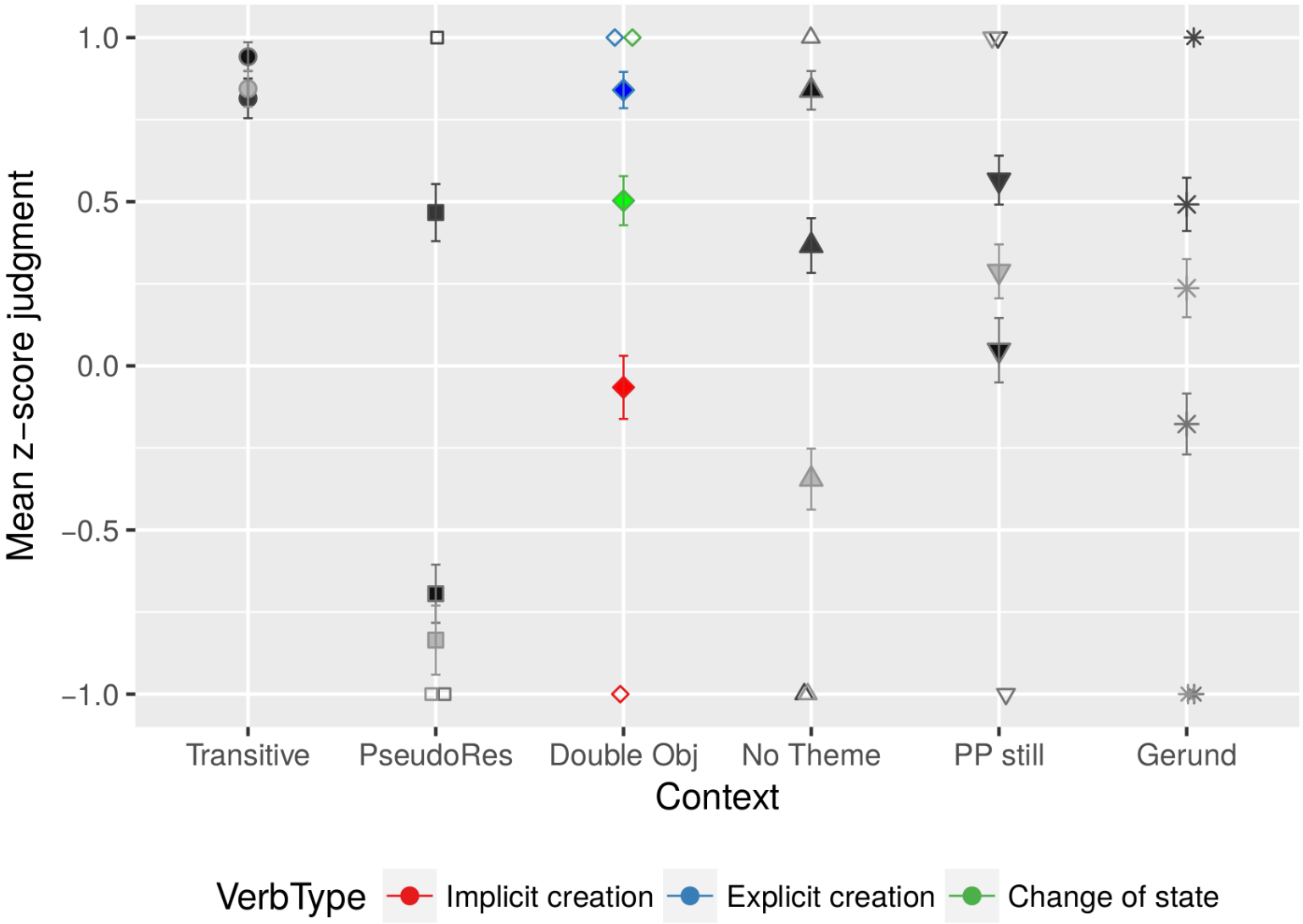
All relevant differences (non-overlapping SEs) significant at $\alpha = 0.05$.

Levinson



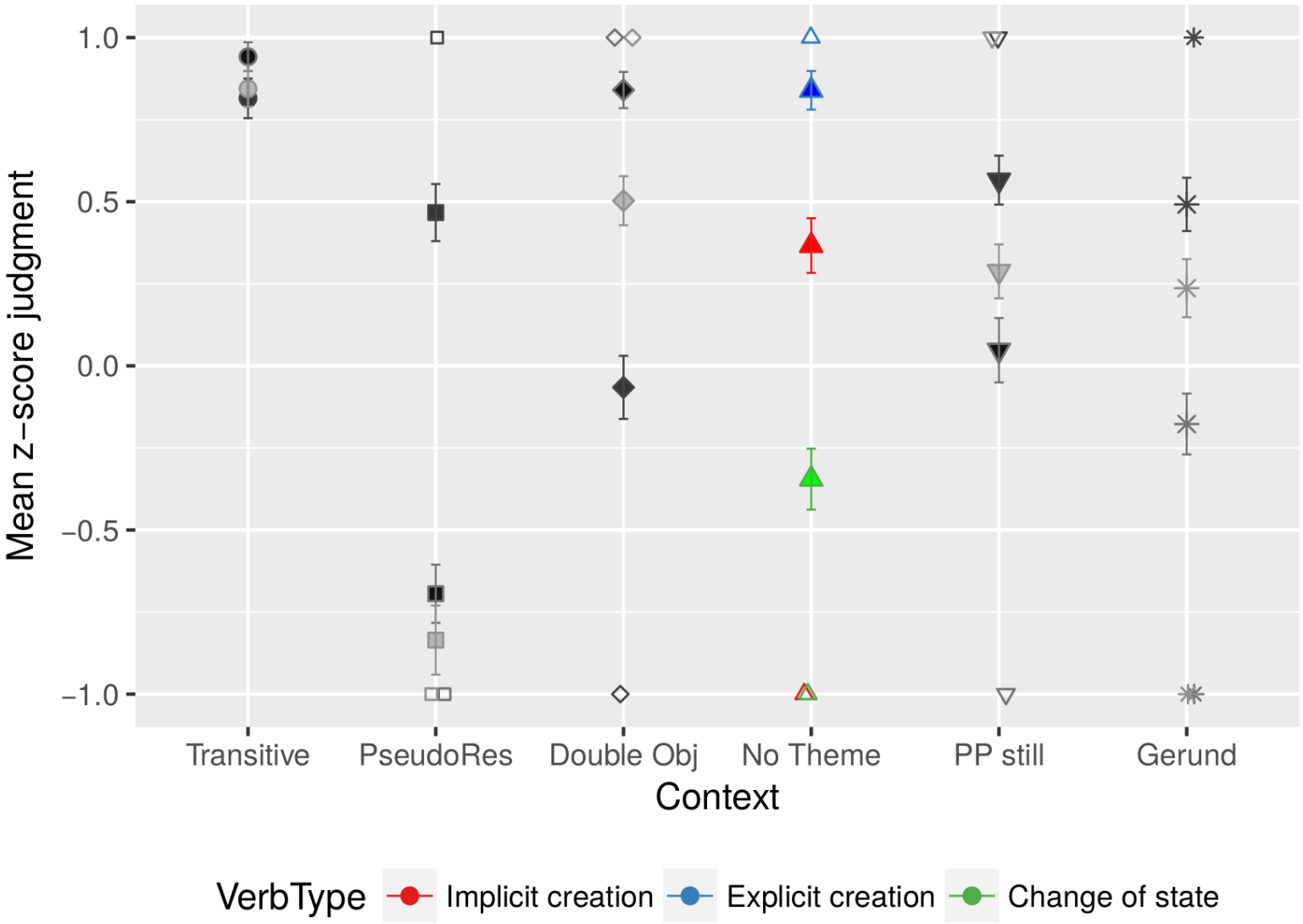
All relevant differences (non-overlapping SEs) significant at $\alpha = 0.05$.

Levinson



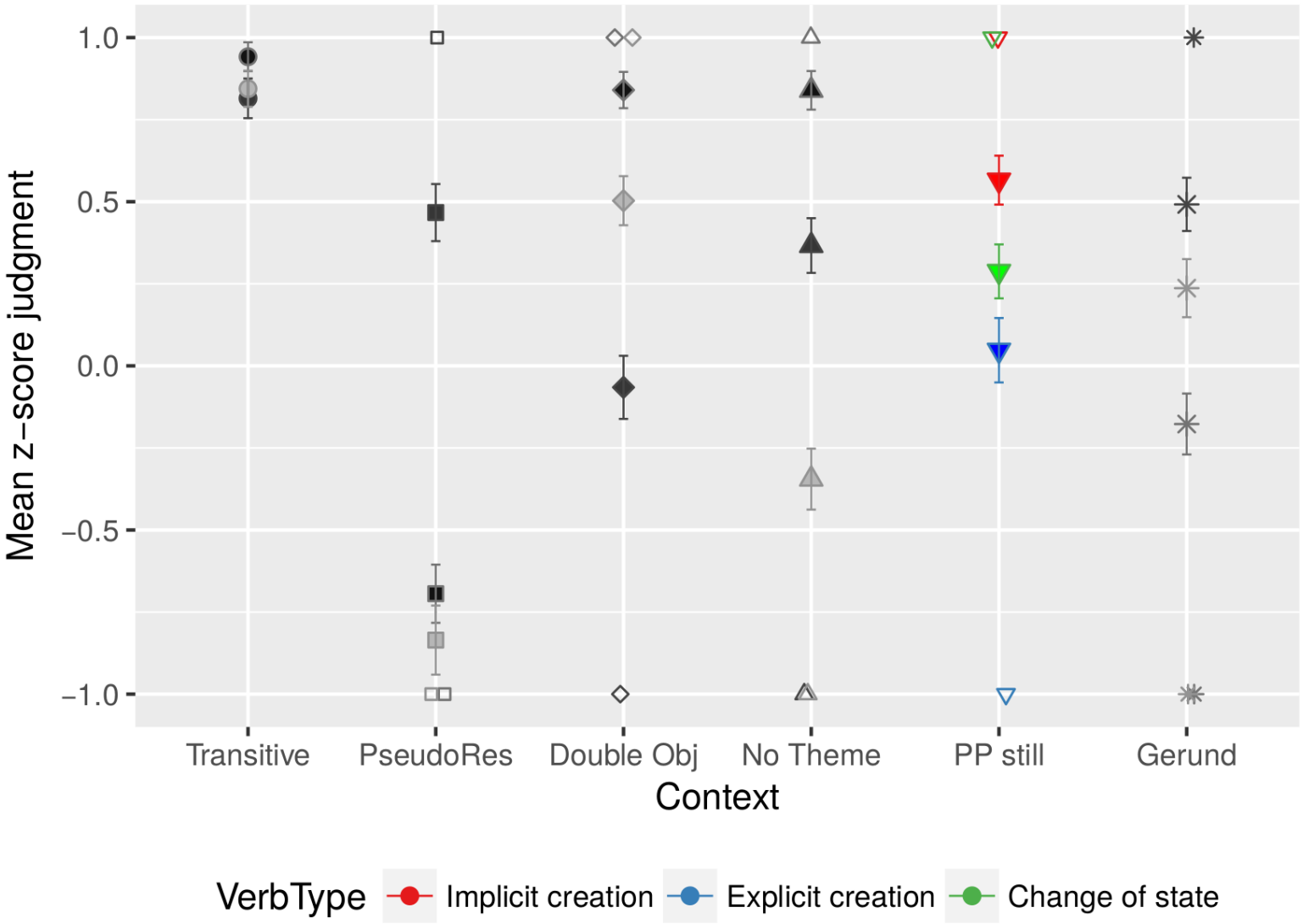
All relevant differences (non-overlapping SEs) significant at $\alpha = 0.05$.

Levinson



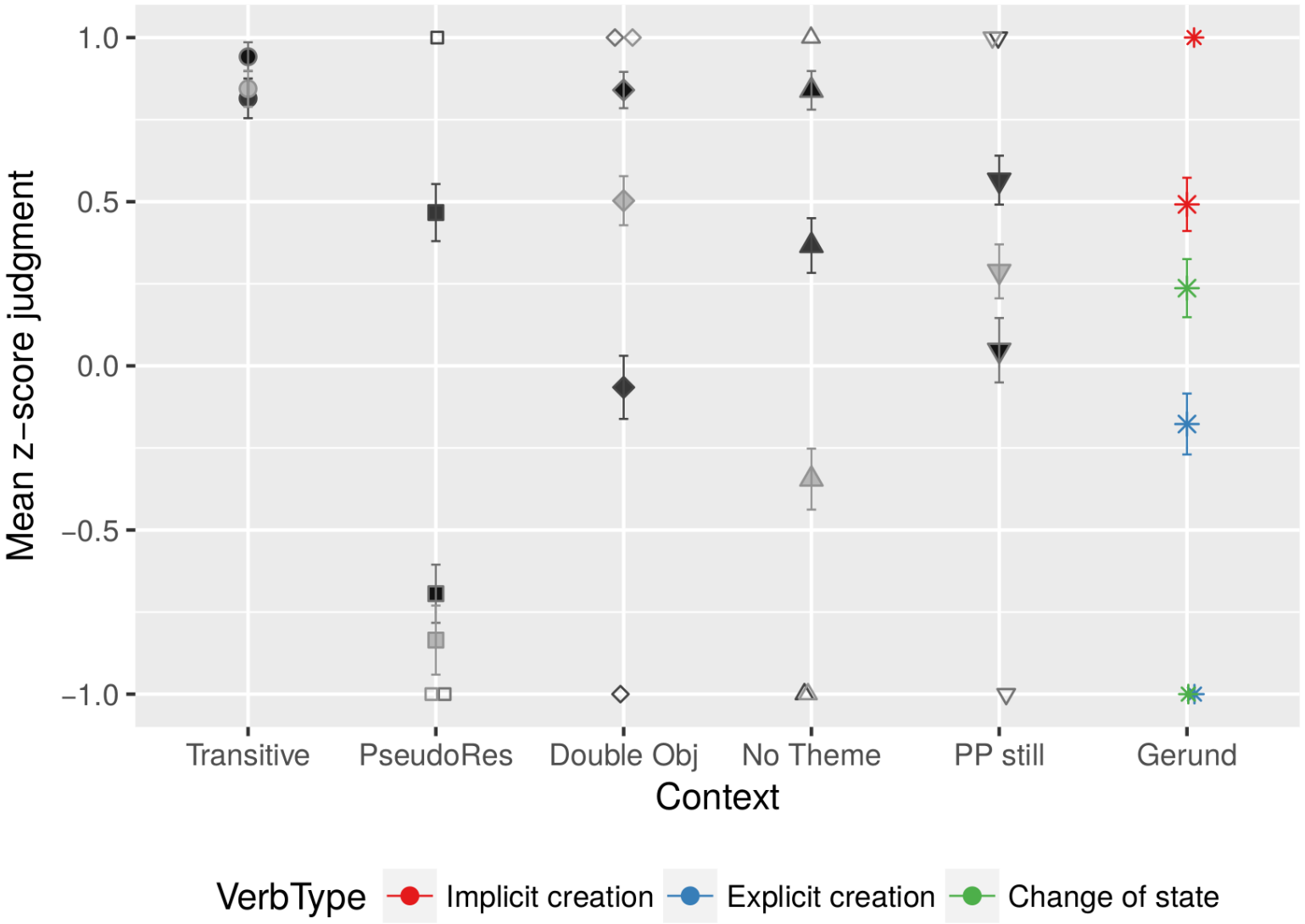
All relevant differences (non-overlapping SEs) significant at $\alpha = 0.05$.

Levinson



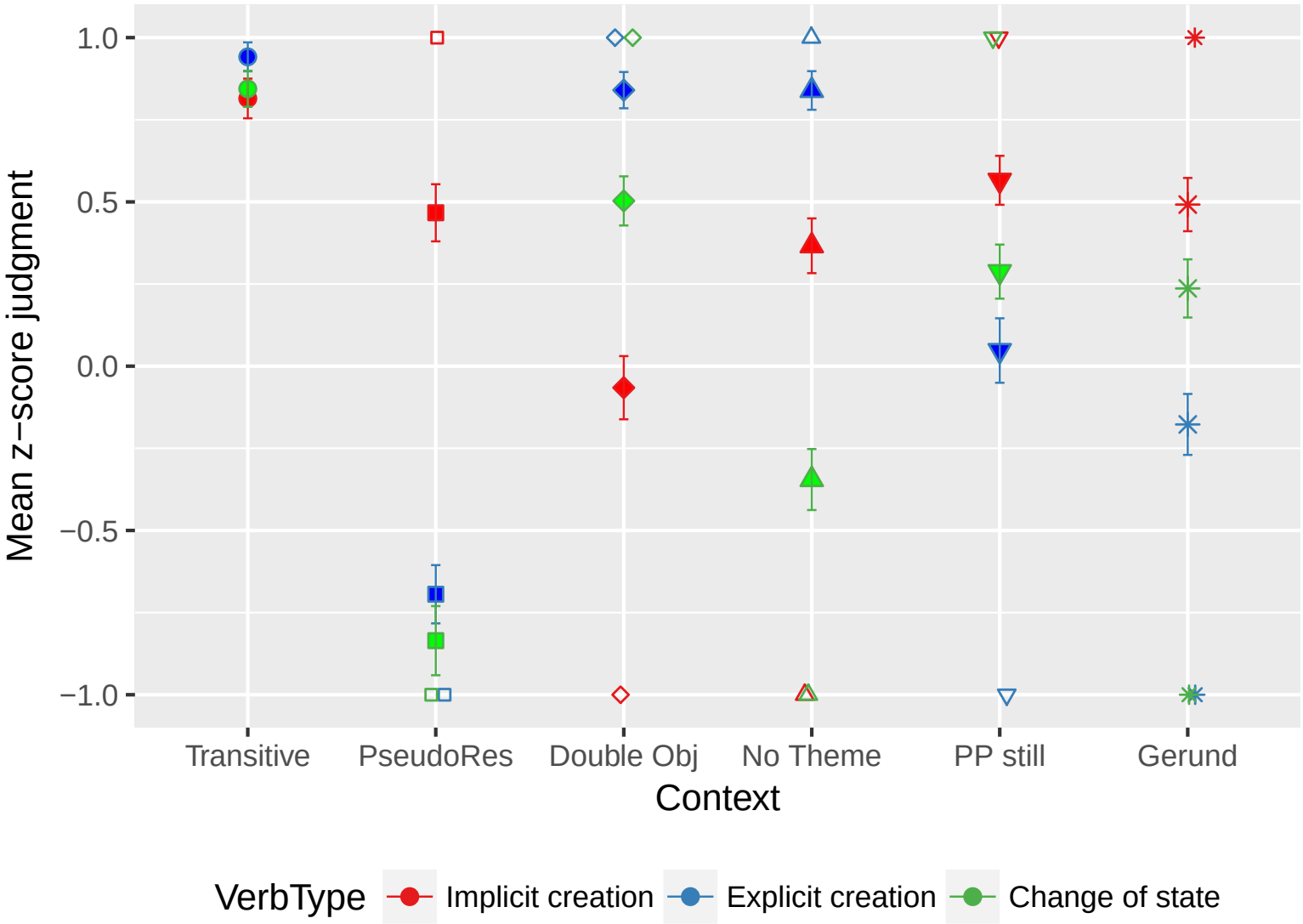
All relevant differences (non-overlapping SEs) significant at $\alpha = 0.05$.

Levinson



All relevant differences (non-overlapping SEs) significant at $\alpha = 0.05$.

Levinson



All relevant differences (non-overlapping SEs) significant at $\alpha = 0.05$.

Levinson

Summary

1. The hypothesized differences are attested robustly.
2. Beyond the striking conformity to the predictions, an unexpected finding was the two conditions which were rated slightly more acceptable than expected.
   - Root-NoTheme.
   - COS-Gerund.

- See Irwin and Kastner (2020) for discussion of these patterns.
- Would also be interesting to look at variation, as discussed yesterday:

(1) a. *John piled his mom some cushions. (3.86 ± 1.70)

b. *?John stacked his uncle some wood. (4.79 ± 1.65)

c. ??John braided his captain some rope. (5.29 ± 1.69)

Next: evaluating another classification.

1. Recap
2. Experimental: Are the patterns robust?
   - Levinson
   - Changes of state
3. Experimental: Do they generalize?
4. Computational: Can they be derived from distributions?
5. Summary

Changes of state

- Earlier this week we briefly mentioned $\sqrt{\text{CRACK}}$, with its additional function `become`.
- Compare that with a root like $\sqrt{\text{COOL}}$:

(2) $[\![\sqrt{\text{COOL}}]\!] = \lambda x \lambda s[\text{cool}(x, s)]$

(3) $[\![\sqrt{\text{CRACK}}]\!] = \lambda x \lambda s[\text{cracked}(x, s) \wedge \exists e[\text{become}(s, e)]]$

- What's special about *crack*?

Changes of state

Two kinds of COS verbs. (Dixon 1982; Spathas and Micheliodakis 2020; Beavers et al. 2021)

Property Concept roots: simple adjectival forms and marked verbal forms.

Property concept roots

- COS verbs.
- De-adjectival.
- Verb $\leftrightarrow$ Adj : *cool/cool(ed)*, *dry/dry(ed)*,
- Adj $\rightarrow$ Verb: *short/shorten(ed)*, *wide/widen(ed)*, *red/redden(ed)*, *hard/harden(ed)*

Result roots: simple verbal forms and marked adjectival forms.

Result roots

- COS verbs.
- Not de-adjectival.
- Verb $\rightarrow$ Adj : *burn/burned, melt/melted, grow/grown, wrinkle/wrinkled, destroy/destroyed.*
- Adj $\rightarrow$ Verb: ✗ (**the bake(en) cake*)

Changes of state

Two kinds of COS verbs. (Dixon 1982; Spathas and Micheliodakis 2020; Beavers et al. 2021)
Property Concept roots: simple adjectival forms and marked verbal forms.

Property concept roots

- COS verbs.
- De-adjectival.
- Verb $\leftrightarrow$ Adj : *cool/cool(ed)*, *dry/dry(ed)*,
- Adj $\rightarrow$ Verb: *short/shorten(ed)*, *wide/widen(ed)*, *red/redden(ed)*, *hard/harden(ed)*

Result roots: simple verbal forms and marked adjectival forms.

Result roots

- COS verbs.
- Not de-adjectival.
- Verb $\rightarrow$ Adj : *burn/burned*, *melt/melted*, *grow/grown*, *wrinkle/wrinkled*, *destroy/destroyed*.
- Adj $\rightarrow$ Verb: ✗ (**the bake(en) cake*)

Changes of state

Semantically, simple PC deverbal adjectives do not entail change:

(4) a. The bright(/#brightened) photo has never brightened.
b. The red(/#reddened) dirt has never reddened.

Result root adjectives entail change:

(5) a. #The cooked chicken has never cooked.
b. #The shattered vase has never shattered.

Changes of state

Also semantically, PC verbs allow restitutive modification:

(6) Jessie flattened the rug again, and it had been flat/flattened before.

Result verbs disallow restitutive modification:

(7) [A store makes their shirts in the back. Jessie buys one and leaves with it, but then decides they don't not want it. They takes the shirt back to exchange it.]
Jessie returned the shirt again.
[put it in the state of being `at-origin`]

Changes of state

Beavers et al. (2021):

1. PC adjectives are simple, result adjectives are complex.
2. Simplex PC adjectives don't entail change, result adjectives always do.
3. PC verbs allow restitutive *again*, result adjectives don't (require repetitive).

$\Rightarrow$ PC adjectives are simple states, result adjectives entail a change leading to that state.

(8) $[\![\sqrt{\text{COOL}}]\!] = \lambda x \lambda s[\text{short}(x, s)]$

(9) $[\![\sqrt{\text{CRACK}}(\text{ed})]\!] = \lambda x \lambda s[\text{cracked}(x, s) \wedge \exists e[\text{become}(s, e)]]$

But is this a quirk of English?

Changes of state

Beavers et al. (2021):

Semantics

Ran the same semantic diagnostics on translational equivalents in Greek (Indo-European), Kakataibo (Panoan), Kinyarwanda (Bantu), Hebrew (Semitic) and Marathi (Indic).

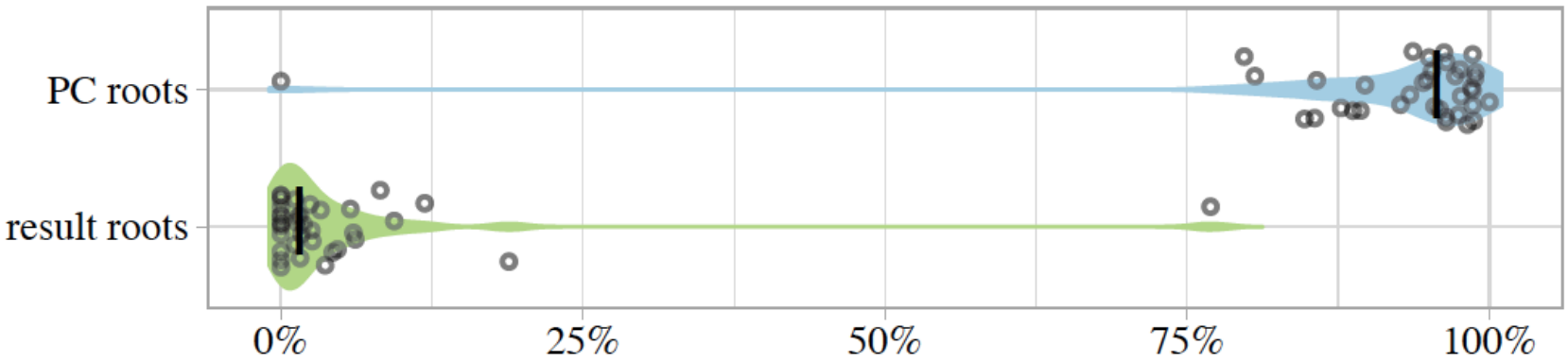
Morphology

Mined dictionaries and grammars for 88 languages, creating mini-paradigms:

Type	root	underlying	simple	inch	caus	result
PC	$\sqrt{\text{RED}}$	—	red	redden	redden	reddened
Res	$\sqrt{\text{SHATTER}}$	—	—	shatter	shatter	shattered

Changes of state

Percentage of languages with simple (underived) statives:



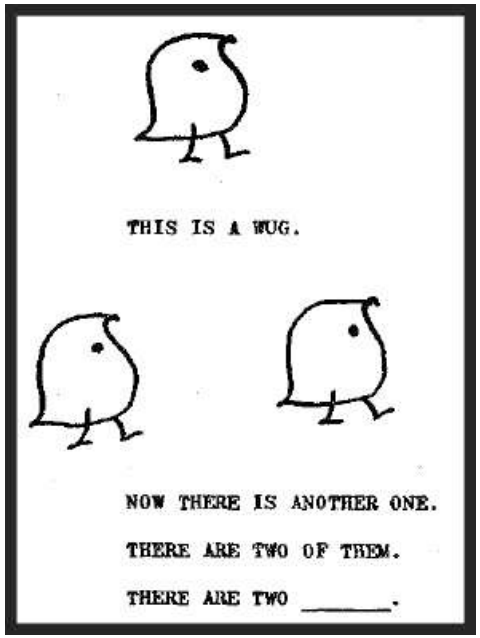
Check out the paper for lots of discussion and follow-up analyses.

- Distinction between two kinds of COS roots.
- Semantic and morphological diagnostics in English.
- Semantic ones replicated qualitatively in a number of languages.
- Morphological ones replicated in a large quantitative sample.

1. Recap
2. Experimental: Are the patterns robust?
3. Experimental: Do they generalize?
   - Learning novel roots
   - Learning novel affixes
4. Computational: Can they be derived from distributions?
5. Summary

Learning novel roots

People are pretty good at learning novel roots/stems.



Berko (1958), Albright and Hayes (2003), and many many others.

But nobody has really probed the lexical semantic distinctions we've been discussing. What would that look like?

Learning novel roots

- Examples (10)–(11) should tell you that the novel verb *to wug* is... Manner or Result?

(10) ✓ Logan wugged and wugged and wugged.
(11) ✓ Logan wugged the blix halfway.

- And then you'd judge (12) accordingly.

(12) ✗ The wind accidentally wugged the blix.

- Manner verbs require animate subjects.

Learning novel roots

- Examples (10)–(11) should tell you that the novel verb *to wug* is... Manner or Result?

(10) ✓Logan wugged and wugged and wugged.
(11) ✓Logan wugged the blix halfway.

- And then you'd judge (12) accordingly.

(12) ✗The wind accidentally wugged the blix.

- Manner verbs require animate subjects.

1. Recap
2. Experimental: Are the patterns robust?
3. Experimental: Do they generalize?
   - Learning novel roots
   - Learning novel affixes
4. Computational: Can they be derived from distributions?
5. Summary

Learning novel affixes

Merkx et al. (2011): are affixes learned through distribution alone or also semantics?

Stem	Type	Example	(Definition)
verb	place	kickort	A large field used by footballers to practise penalties
noun	place	cointund	The factory in which the twenty pence coin is produced
verb	tool	pourlabe	A bottle cap used for pouring exact measures
noun	tool	wheathoke	A harvesting tool used by farmers in the Middle Ages
verb	person	sleepnept	A participant in a study about the effects of napping
noun	person	rugete	A person who imports and sells handmade carpets
verb	cost	leapesh	The cost of having a stuntman jump out of a building
noun	cost	bombaph	The cost of buying enough explosives to blow up a car

- Participants managed to learn all the new affixes.
- It was hard for them to reject *sleephoke* even though they never saw it – meaning they generalized the affix.
- Semantics helped further.

Learning novel affixes

Merkx et al. (2011): are affixes learned through distribution alone or also semantics?

Stem	Type	Example	(Definition)
verb	place	kickort	A large field used by footballers to practise penalties
noun	place	cointund	The factory in which the twenty pence coin is produced
verb	tool	pourlabe	A bottle cap used for pouring exact measures
noun	tool	wheathoke	A harvesting tool used by farmers in the Middle Ages
verb	person	sleepnept	A participant in a study about the effects of napping
noun	person	rugete	A person who imports and sells handmade carpets
verb	cost	leapesh	The cost of having a stuntman jump out of a building
noun	cost	bombaph	The cost of buying enough explosives to blow up a car

- Participants managed to learn all the new affixes.
- It was hard for them to reject *sleephoke* even though they never saw it – meaning they generalized the affix.
- Semantics helped further.

Learning novel affixes

Tamminen et al. (2015): Does the meaning of the affix matter?

Affix	Examples of trained novel words and associated meanings
-nule	Bricknule is the labourer who operates the oven which hardens clay to brick
	Foxnule is someone who looks after a fox harmed in a car accident
-afe	Crabafe is the zoo building where you can see exotic crab species
	Gunafe is the section of an armoury where one can find a gun
-lomb	Fetchlomb is an extendable arm used to fetch small items without getting up
	Mowlomb is a popular machine which can mow the lawn automatically
-esh	Warnesh is the yearly cost the state pays to warn expatriates of danger
	Begesh is the amount children pay to gangsters to be allowed to beg

Learning novel affixes

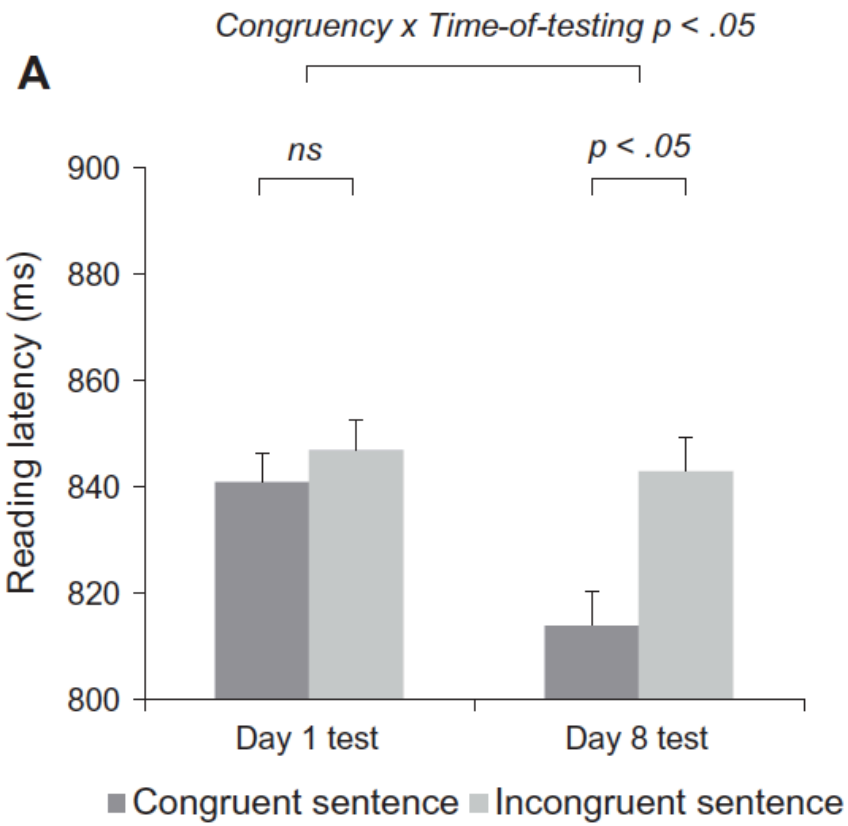
Congruency test:

Condition	Sentence	Affix
Congruent	It was an honour to be visited by the sandnule	Person
	The man rushed to get inside the beanafe	Place
	They were taught how to operate a warmlomb	Tool
	He thought it would help if he paid them a hurlesh	Cost
Incongruent	The company had just relocated to a peachnule	Person
	The manager often argued with a pigafe	Place
	They were arrested for not paying the required rentlomb	Tool
	They always made fun of her for using a readesh	Cost

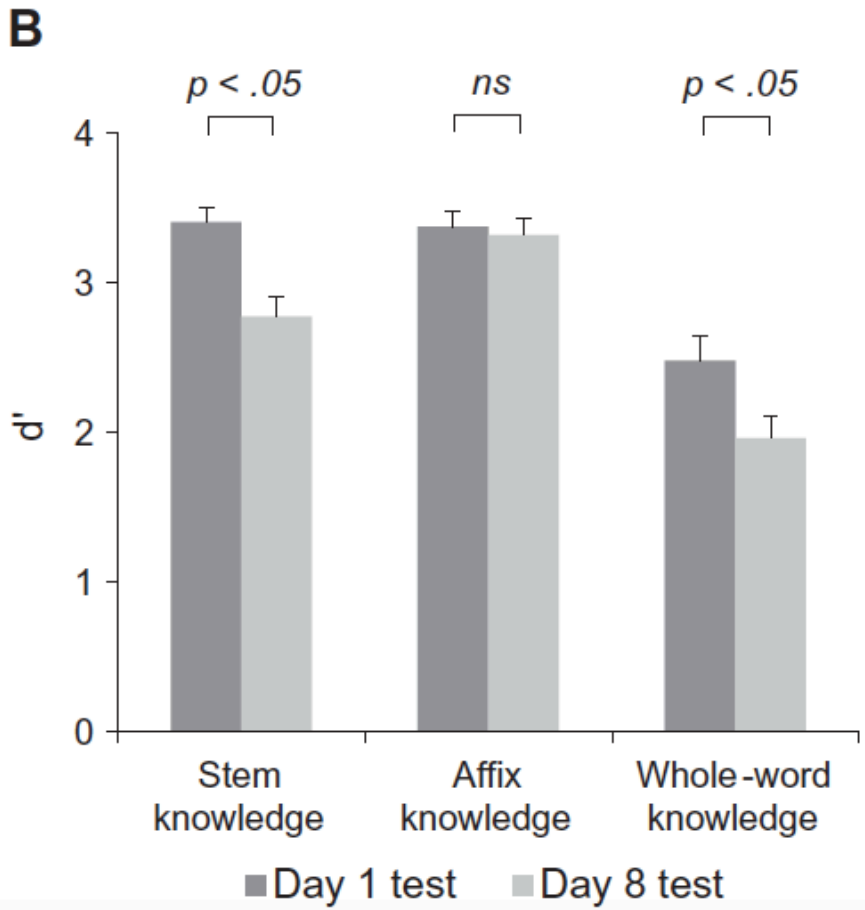
Learning novel affixes

Common assumption: Lower response time = less processing (easier task).

Participants learned the meanings:



And learned affixes, not whole words (and also didn't care about stems as much)



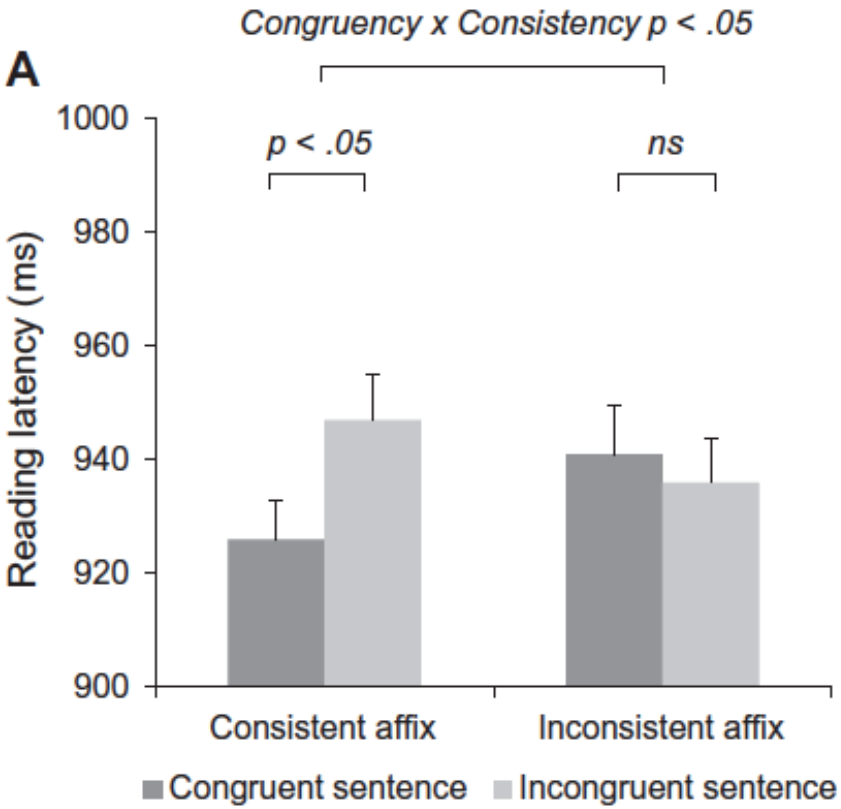
Learning novel affixes

Does **consistency** in meaning matter?

Semantic consistency	Novel word (meaning category)
Consistent	Buildnule (cost), Sleepnule (cost)
	Bringane (place), Lockane (place)
	Crewose (tool), Bombose (tool)
	Girltege (person), Graintege (person)
Inconsistent	Knitlomb (person), Swimlomb (tool)
	Creepesh (place), Grabesh (cost)
	Hairuck (cost), Gunuck (person)
	Sheephalk (tool), Creamhalk (place)

Learning novel affixes

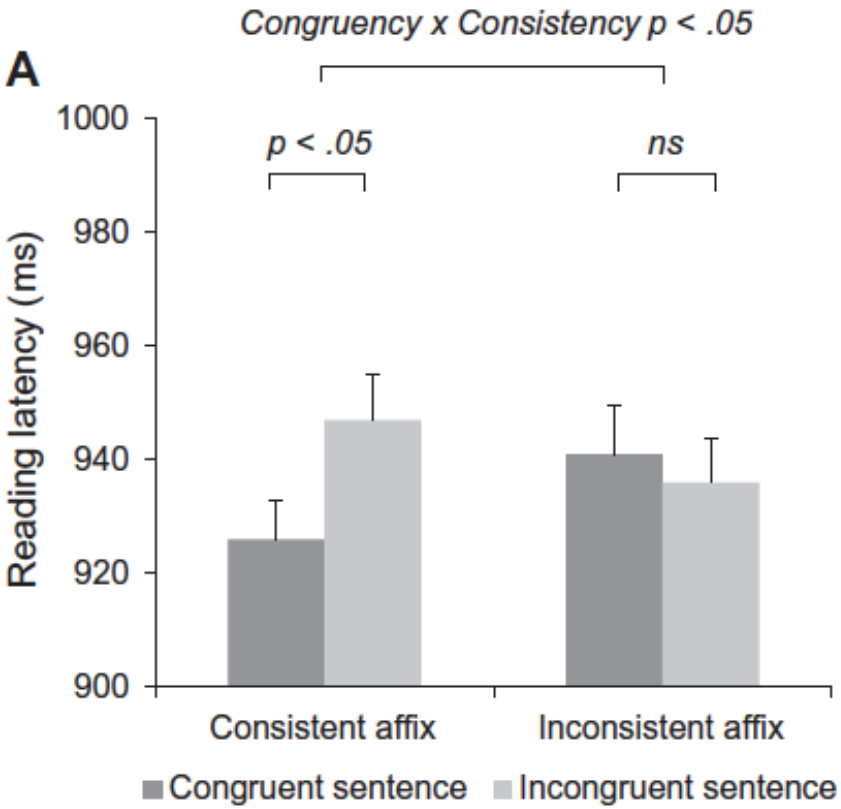
Consistency matters:



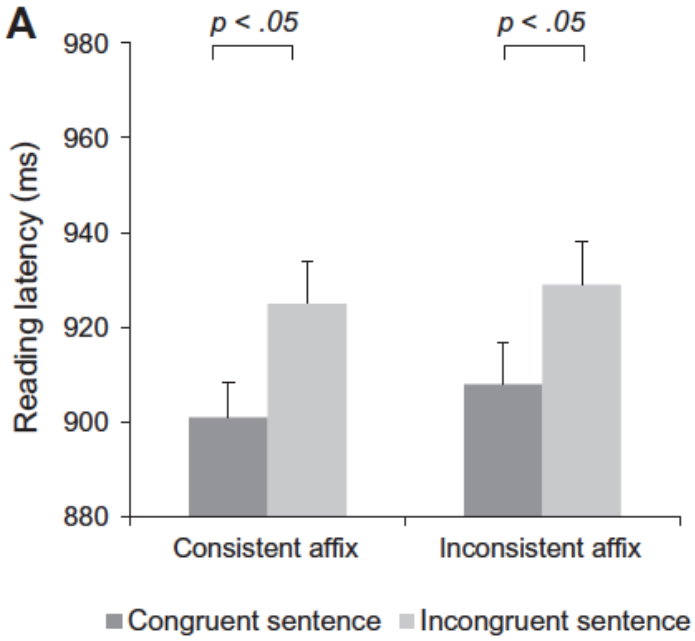
But it was ok again if they first learned one meaning, had 24 hours to consolidate, then learned the other.

Learning novel affixes

Consistency matters:



But it was ok again if they first learned one meaning, had 24 hours to consolidate, then learned the other.



Learning novel affixes

Summary

- People can learn novel affixes.
- If you put enough work into:
  - The materials.
  - The task.
  - The evaluation techniques.
- A consolidation period helps.
- Separating meanings in training helps.
- Some questions:
  1. Does it matter that these are "derivational" affixes?
  2. Do some meanings combine better with some stems?
  3. How would children do on this task?

$\Rightarrow$ Lots still to do!

Learning novel affixes

Summary

- People can learn novel affixes.
- If you put enough work into:
  - The materials.
  - The task.
  - The evaluation techniques.
- A consolidation period helps.
- Separating meanings in training helps.
- Some questions:
  1. Does it matter that these are "derivational" affixes?
  2. Do some meanings combine better with some stems?
  3. How would children do on this task?

$\Rightarrow$ Lots still to do!

1. Recap
2. Experimental: Are the patterns robust?
3. Experimental: Do they generalize?
4. Computational: Can they be derived from distributions?
   - Word embeddings
   - Experiment 1: Manner/Result in English
   - Experiment 2: Manner/Result in Hebrew
5. Summary

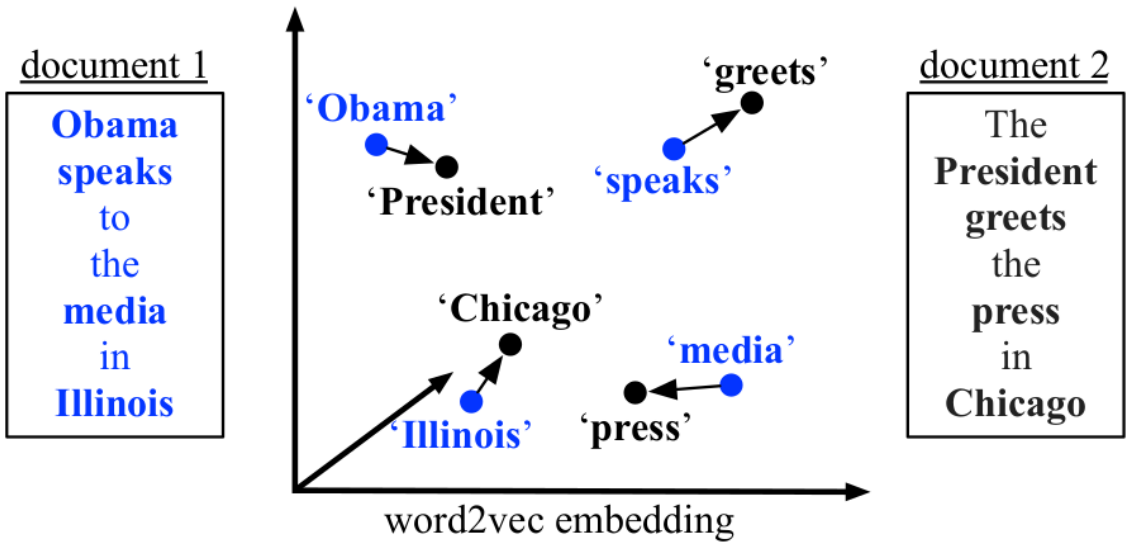
Word embeddings

Time	Language and Logic (LaLo)		Language and Computation (LaCo)		Logic and Computation (LoCo)		Workshops
9.00am - 10.30am	*Yoad Winter*: Formal Semantics of Natural Language (Foundational) Room: TBA	*Paul Dekker*: Outline of a Theory of Interpretation (Advanced) Room: TBA	*Diego Frassinelli and Sabine Schulte Im Walde*: Cognitive and Computational Models of Abstractness (Introductory) Room: TBA	*Lasha Abzianidze*: Natural Language Reasoning with a Natural Theorem Prover (Advanced) Room: TBA	*Aleks Knoks*: Defeasible Logics with Applications to Normative Systems and Philosophy (Foundational) Room: TBA	*Hans van Ditmarsch and Malvin Gattinger*: Knowledge and Gossip (Advanced) Room: TBA	*Timothée Bernard and Grégoire Winterstein*: Bridges and Gaps between Formal and Computational Linguistics Room: TBA
10.30am - 11.00am	Coffee Break						
11.00am - 12.30pm	*Itamar Kastner*: The Logical Form of Lexical Semantics (Introductory) Room: TBA	*Jeremy Goodman and Cian Dorr*: Theory-Building in Higher Order Languages (Advanced) Room: TBA	*Stephanie Evert and Gabriella Lapesa*: Hands-on Distributional Semantics for Linguistics using R (Foundational) Room: TBA	*Gijs Wijnholds and Michael Moortgat*: Compositional Models of Vector-based Semantics: From Theory to Tractable Implementation (Advanced) Room: TBA	*Thomas Icard and Krzysztof Mierzewski*: *Logic & Probability* (Introductory) Room: TBA		
12.30pm - 2.00pm	Lunch						
2.00pm - 3.30pm	*David Boylan and Matthew Mandelkern*: Conditionals and Information-Sensitivity (Introductory) Room: TBA	*Timothée Bernard and Justin Bledin*: Negative Events and Truthmaker Semantics (Advanced) Room: TBA	*Gabriella Lapesa and Eva Maria Vecchi*: Argument Mining between NLP and Social Sciences (Introductory) Room: TBA	*David Traum*: Computational Models of Grounding in Dialogue (Advanced) Room: TBA	*Luca Reggio and Tomáš Jakl*: Relating Structure to Power: An Invitation to Game Comonads (Advanced) Room: TBA **Sponsored by EACSL**		*Jelka van der Sluis and James Pustejovsky*: Annotation, Recognition and Evaluation of Actions II (AREA-II) *(Tuesday, Wednesday and Thursday only)* Room: TBA
3.30pm - 3.50pm	Coffee Break						

And that's just Week 1!

Word embeddings

- Create an abstract representation of words in a corpus (vector space).
- Calculate co-occurrence of words and "contexts" (other words).

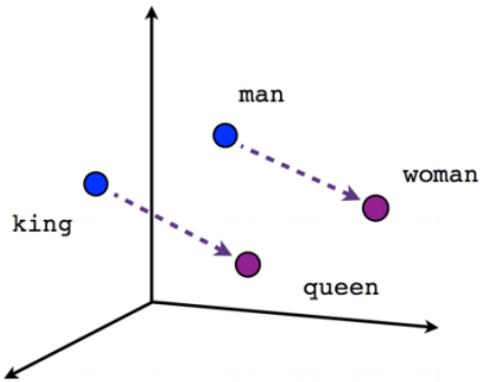


- We get an abstract, numerical representation of each word: a vector.
  dog = [2.972568, -0.76399034, 1.3605528, -2.036042, -2.3865438, ...]
- From Latent Semantic Analysis (Landauer and Dumais 1997) to neural networks.
  - `word2vec` (Mikolov et al. 2013), `GloVe` (Pennington et al. 2014).
  - `ELMo` (Peters et al. 2018), `BERT` (Devlin et al. 2019) and other language models.

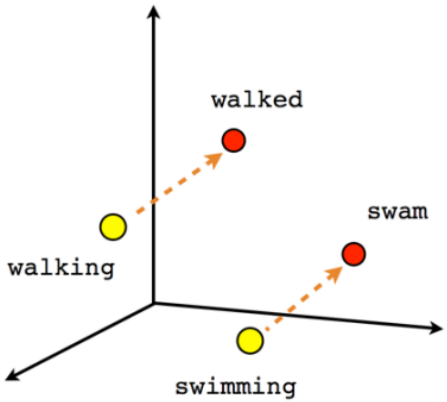
Word embeddings

Allow for computations such as:

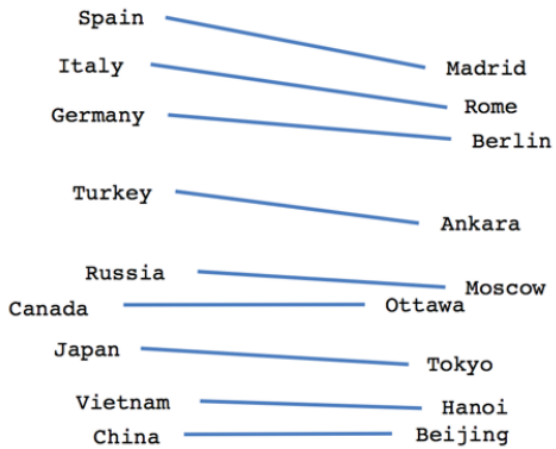
- Similarity: **closest things** to *dog* are:
  - ('cat', 0.888), ('dog,', 0.868), ('rabbit', 0.824), ('fox', 0.809), ('puppy', 0.789), ('dogs', 0.783), ('horse', 0.777), ('pet', 0.775), ('cat,', 0.768), ('kitten', 0.765)
- Analogies / algebra
  - Ireland : Dublin :: Scotland : ?
  - Man − Woman + King = ?



Male-Female



Verb tense



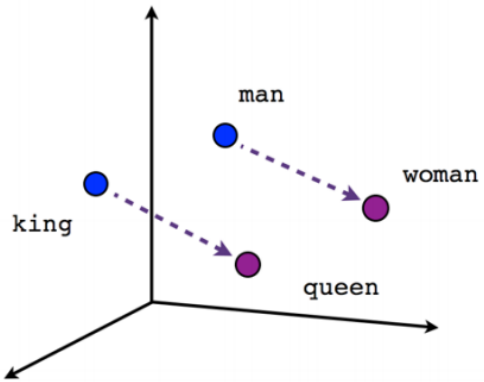
Country-Capital

(But cf. Linzen 2016 and tons of biases)

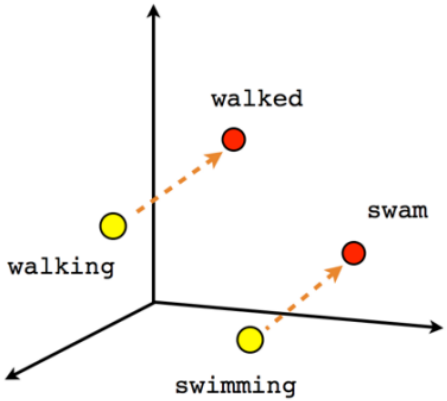
Word embeddings

Allow for computations such as:

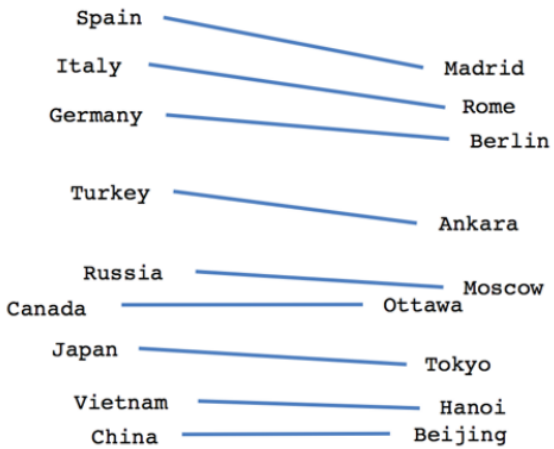
- Similarity: **closest things** to *dog* are:
  - ('cat', 0.888), ('dog,', 0.868), ('rabbit', 0.824), ('fox', 0.809), ('puppy', 0.789), ('dogs', 0.783), ('horse', 0.777), ('pet', 0.775), ('cat,', 0.768), ('kitten', 0.765)
- Analogies / algebra
  - Ireland : Dublin :: Scotland : **Edinburgh**
  - Man – Woman + King = ?



Male-Female



Verb tense



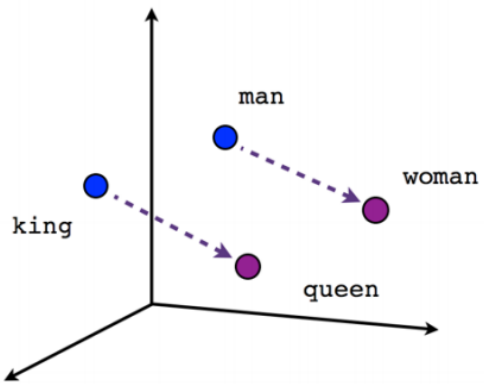
Country-Capital

(But cf. Linzen 2016 and tons of biases)

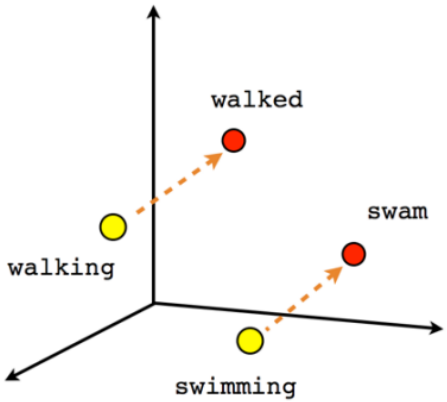
Word embeddings

Allow for computations such as:

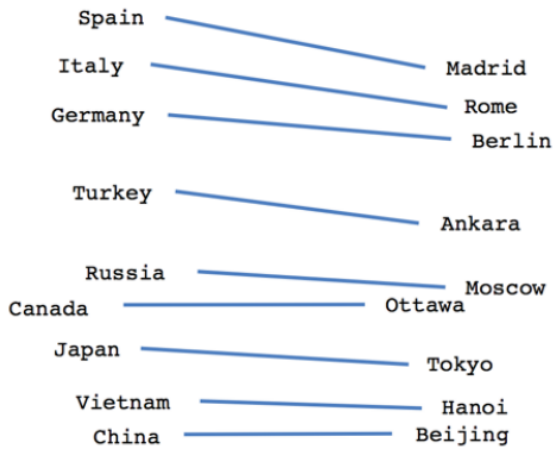
- Similarity: **closest things** to *dog* are:
  - ('cat', 0.888), ('dog,', 0.868), ('rabbit', 0.824), ('fox', 0.809), ('puppy', 0.789), ('dogs', 0.783), ('horse', 0.777), ('pet', 0.775), ('cat,', 0.768), ('kitten', 0.765)
- Analogies / algebra
  - Ireland : Dublin :: Scotland : **Edinburgh**
  - Man – Woman + King = **Queen**



Male-Female



Verb tense



Country-Capital

(But cf. Linzen 2016 and tons of biases)

Word embeddings

- There is lots to say about these and their relationship with “real” semantics.

Time	Language and Logic	(LaLo)	Language and Computation	(LaCo)	Logic and Computation	(LoCo)	Workshops
9.00am - 10.30am	*Yoad Winter:* Formal Semantics of Natural Language (Foundational) Room: TBA	*Paul Dekker:* Outline of a Theory of Interpretation (Advanced) Room: TBA	*Diego Frassinelli and Sabine Schulte Im Walde:* Cognitive and Computational Models of Abstractness (Introductory) Room: TBA	*Lasha Abzianidze:* Natural Language Reasoning with a Natural Theorem Prover (Advanced) Room: TBA	*Aleks Knoks:* Defeasible Logics with Applications to Normative Systems and Philosophy (Foundational) Room: TBA	*Hans van Ditmarsch and Malvin Gattinger:* Knowledge and Gossip (Advanced) Room: TBA	*Timothée Bernard and Grégoire Winterstein:* Bridges and Gaps between Formal and Computational Linguistics Room: TBA
10.30am - 11.00am	Coffee Break						
11.00am - 12.30pm	*Itamar Kastner:* The Logical Form of Lexical Semantics (Introductory) Room: TBA	*Jeremy Goodman and Cian Dorr:* Theory-Building in Higher Order Languages (Advanced) Room: TBA	*Stephanie Evert and Gabriella Lapesa:* Hands-on Distributional Semantics for Linguistics using R (Foundational) Room: TBA	*Gijs Wijnholds and Michael Moortgat:* Compositional Models of Vector-based Semantics: From Theory to Tractable Implementation (Advanced) Room: TBA	*Thomas Icard and Krzysztof Mierzewski:* Logic & Probability (Introductory) Room: TBA		
12.30pm - 2.00pm	Lunch						
2.00pm - 3.30pm	*David Boylan and Matthew Mandelkern:* Conditionals and Information-Sensitivity (Introductory) Room: TBA	*Timothée Bernard and Justin Bledin:* Negative Events and Truthmaker Semantics (Advanced) Room: TBA	*Gabriella Lapesa and Eva Maria Vecchi:* Argument Mining between NLP and Social Sciences (Introductory) Room: TBA	*David Traum:* Computational Models of Grounding in Dialogue (Advanced) Room: TBA	*Luca Reggio and Tomáš Jakl:* Relating Structure to Power: An Invitation to Game Comonads (Advanced) Room: TBA **Sponsored by EACSL**		*Jelka van der Sluis and James Pustejovsky:* Annotation, Recognition and Evaluation of Actions II (AREA-II) *(Tuesday, Wednesday and Thursday only)* Room: TBA
3.30pm - 3.50pm	Coffee Break						

- NLP researchers are really interested in the syntactic capabilities of contemporary language models. (Bowman et al. 2015; Hewitt and Manning 2019; Tenney et al. 2019; Ettinger 2020; Linzen and Baroni 2020; Kogkalidis and Wijnholds 2022)
- We’re essentially already assuming that they learn semantics.
- Talk to many of the people at ESSLLI who know more about these models than me, or see the “further reading” section at the end.

1. Recap
2. Experimental: Are the patterns robust?
3. Experimental: Do they generalize?
4. Computational: Can they be derived from distributions?
   - Word embeddings
   - Experiment 1: Manner/Result in English
   - Experiment 2: Manner/Result in Hebrew
5. Summary

Experiment 1: Manner/Result in English

1. Can word embeddings learn Manner/Result Complementarity?
2. Feed English Wikipedia to a word embedding model.
3. See if Manner and Result cluster differently.

Experiment 1: Manner/Result in English

For the items:

- Used the existing examples in Rappaport Hovav and Levin (2010) and Rappaport Hovav (2017).
- Total of 28 Manner verbs and 29 Result verbs.

For the corpus:

- English Wikipedia (2013).
- Used the full corpus (not lemmatized).
- 5,351 documents, 846M tokens, average word length 6.2 characters.

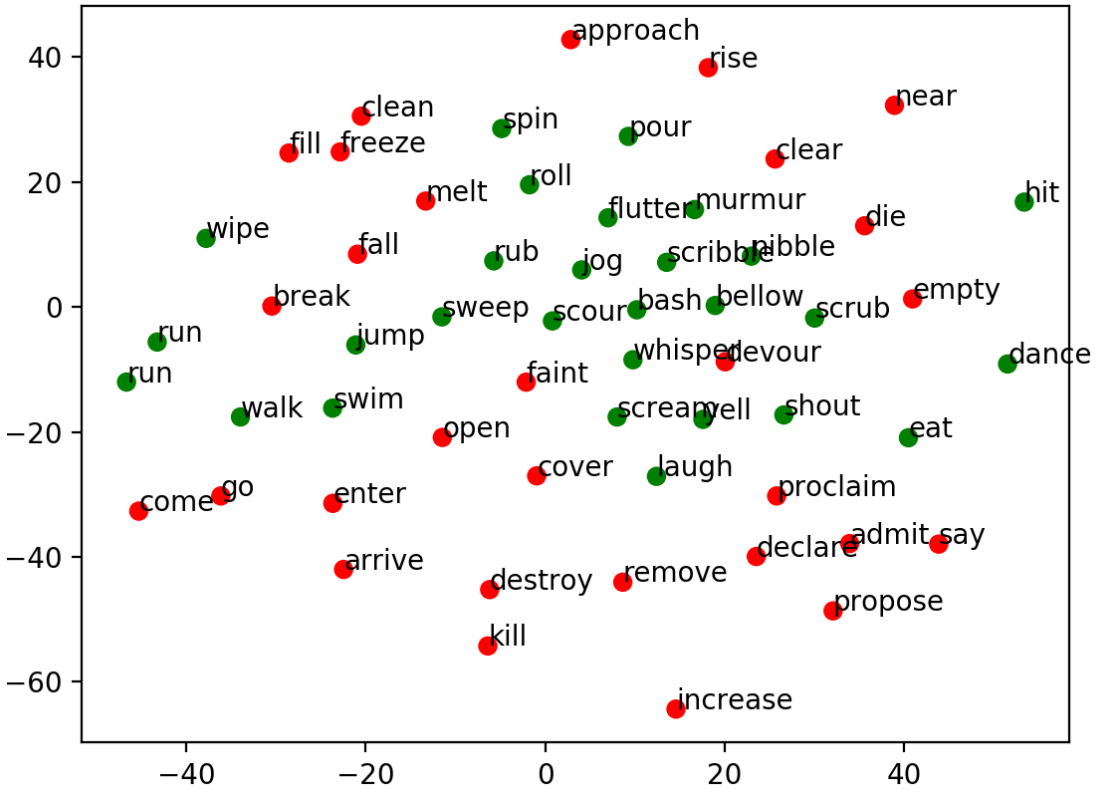
Experiment 1: Manner/Result in English

bash	*murmur*	*scrub*	*admit*	*devour*	*kill*
bellow	*nibble*	*shout*	*approach*	*die*	*melt*
dance	*pour*	*spin*	*arrive*	*empty*	*near*
eat	*roll*	*sweep*	*break*	*enter*	*open*
flutter	*rub*	*swim*	*clean*	*faint*	*proclaim*
hit	*run*	*walk*	*clear*	*fall*	*propose*
jog	*scour*	*whisper*	*come*	*fill*	*remove*
jump	*scream*	*wipe*	*cover*	*freeze*	*rise*
laugh	*scribble*	*yell*	*declare*	*go*	*say*
murmur			*destroy*	*increase*	

Experiment 1: Manner/Result in English

1. Constructed the embeddings using the `word2vec` implementation (Mikolov et al. 2013) in Gensim (Řehůřek and Sojka 2010).
2. Reduced the dimensionalty to two dimensions using t-SNE (van der Maaten and Hinton 2008).
3. Just plotting the data at this point, not even at statistical analysis yet.

Experiment 1: Manner/Result in English



- Not bad.
- You can see that semantically similar words cluster together (but not always).
- *Walk-swim-jump*, *flutter-murmer*, *bellow-whisper-faint*.
- Not the cleanest separation between the two clusters.
- But there seems to be something there.

Experiment 1: Manner/Result in English

1. Levy and Goldberg (2014): use syntactic dependencies for word embeddings.
2. It's what the learner would do anyway.

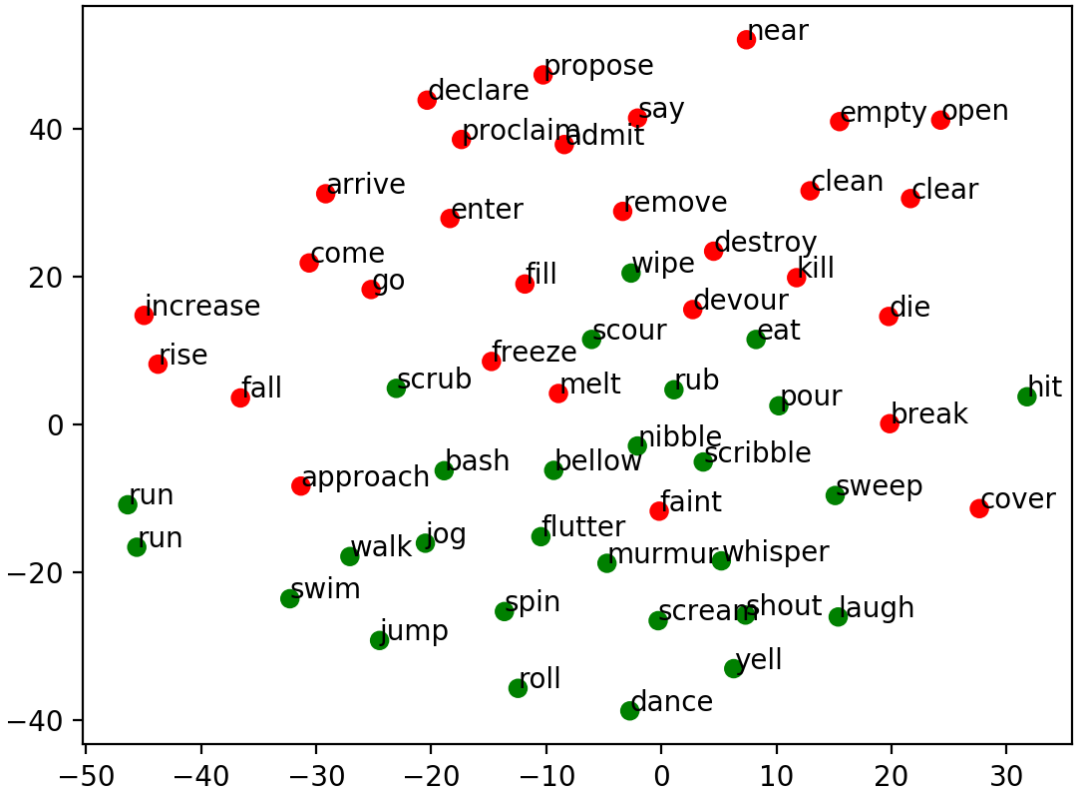
(Landau and Gleitman 1985; Gleitman 1990; Gillette et al. 1999; Fisher et al. 2010)

3. Adding syntax helps in two ways. First, performance is better overall.
4. Second, this kind of model arguably looks for *similarity* rather than *association*.

	Bag of words	Dependencies
Hogwarts	Dumbledore hallows	Sunnydale Greendale
Turing	non-deterministic finite-state	Hamming Hotelling
Florida	Gainesville fla	Texas California

Experiment 1: Manner/Result in English

- Used the English dependency-based embeddings from Levy and Goldberg (2014).
- Running the same analysis, except with embeddings based on dependencies rather than (just) words:



- Looks like improvement.
- *Increase-rise, walk-swim-jump, whisper-murmur-flutter-faint-bellow.*

Experiment 1: Manner/Result in English

1. The cluster is a good start for word-based embeddings.
2. Syntactic dependencies clearly help.
3. Potential for useful error analysis (*faint*).
4. Quantitative evaluation still necessary (e.g. k-fold cross-validation).
5. Can a clustering algorithm learn how to tell the two classes apart in an unsupervised manner?

➤ Can we replicate this in Hebrew?

1. Recap
2. Experimental: Are the patterns robust?
3. Experimental: Do they generalize?
4. Computational: Can they be derived from distributions?
   - Word embeddings
   - Experiment 1: Manner/Result in English
   - Experiment 2: Manner/Result in Hebrew
5. Summary

Experiment 2: Manner/Result in Hebrew

Much-debated question: what is The Semantics of a Semitic root?

(Aronoff 1994; Harley 2014; Kastner 2020b)

		niXYaZ		*XaYaZ*		*heXYiZ*	
a.	$\sqrt{\text{KTB}}$	*nixtav*	'was written'	*katav*	'wrote'	*hextiv*	'dictated'
b.	$\sqrt{\text{KRJ}}$	*nikra*	'was read'	*kara*	'read'	*hekri*	'read out'
c.	$\sqrt{\text{SGR}}$	*nisgar*	'was closed'	*sagar*	'closed'	*hesgir*	'extradited'

Or, looking at nouns in $\sqrt{\text{KB}\int}$:

	kviʃ	'road'
	kibuʃ	'occupation'
(*melafefon*)	*kavuʃ*	'pickle'

Linguists would have good use for a tool quantifying similarity of words.

Experiment 2: Manner/Result in Hebrew

Much-debated question: what is The Semantics of a Semitic root?

(Aronoff 1994; Harley 2014; Kastner 2020b)

		niXYaZ		*XaYaZ*		*heXYiZ*	
a.	$\sqrt{\text{KTB}}$	*nixtav*	'was written'	*katav*	'wrote'	*hextiv*	'dictated'
b.	$\sqrt{\text{KRJ}}$	*nikra*	'was read'	*kara*	'read'	*hekri*	'read out'
c.	$\sqrt{\text{SGR}}$	*nisgar*	'was closed'	*sagar*	'closed'	*hesgir*	'extradited'

Or, looking at nouns in $\sqrt{\text{KB}\int}$:

	kviʃ	'road'
	kibuʃ	'occupation'
(*melafefon*)	*kavuʃ*	'pickle'

Linguists would have good use for a tool quantifying similarity of words.

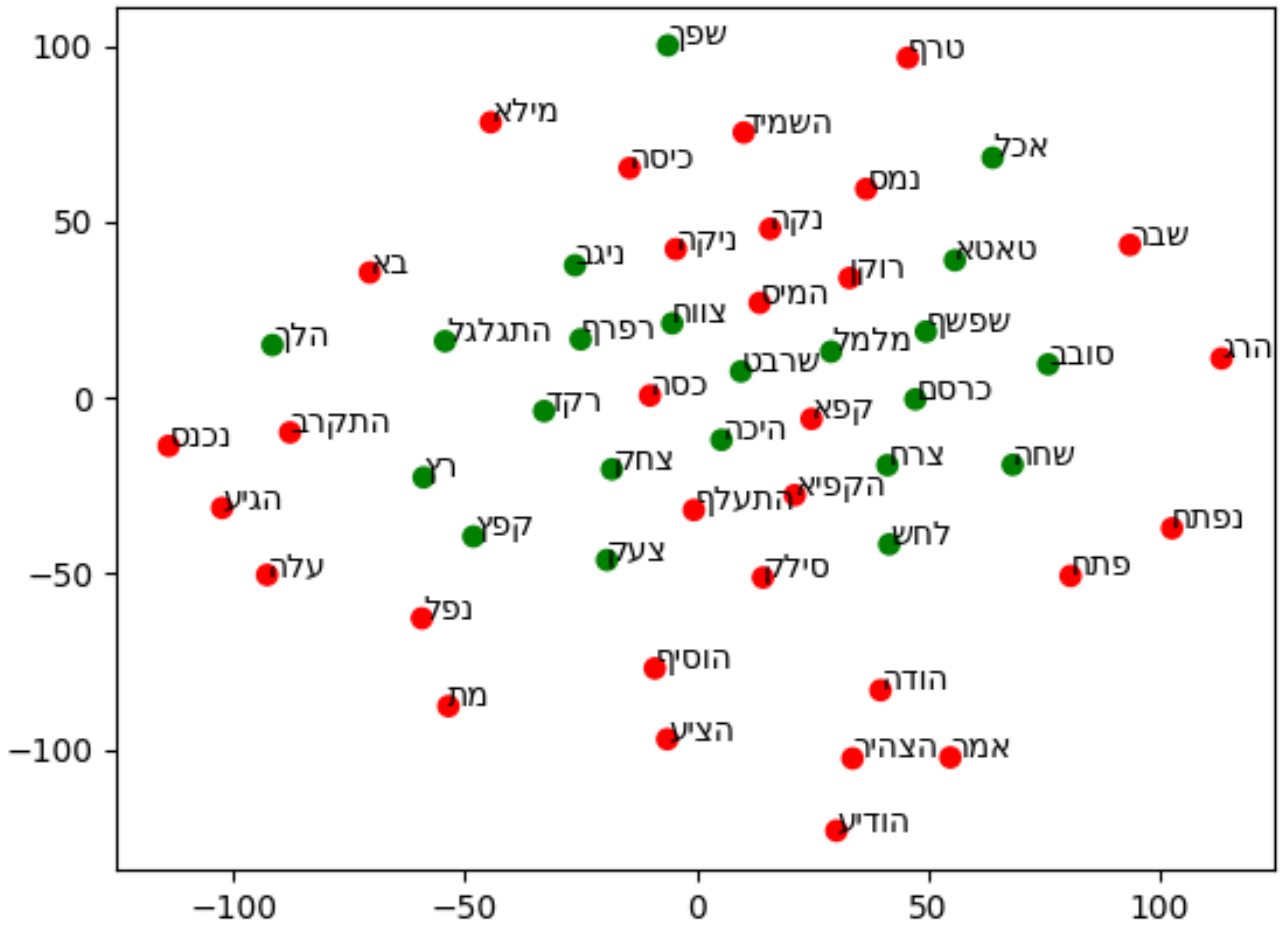
Experiment 2: Manner/Result in Hebrew

First of all, let's replicate the English finding for Manner/Result in Hebrew.

Experiment 2a:

- Translated the English list.
- Ended up with 28 Manner verbs and 31 Result verbs.
  - Added alternations, e.g. *open*: *niftax, patax.*
- Used only the citation forms (3SG.M past).
- Again used Wikipedia (not lemmatized).
- 891 documents, 81M tokens, average word length 10.4 characters.
  - (an order of magnitude smaller than the English corpus)

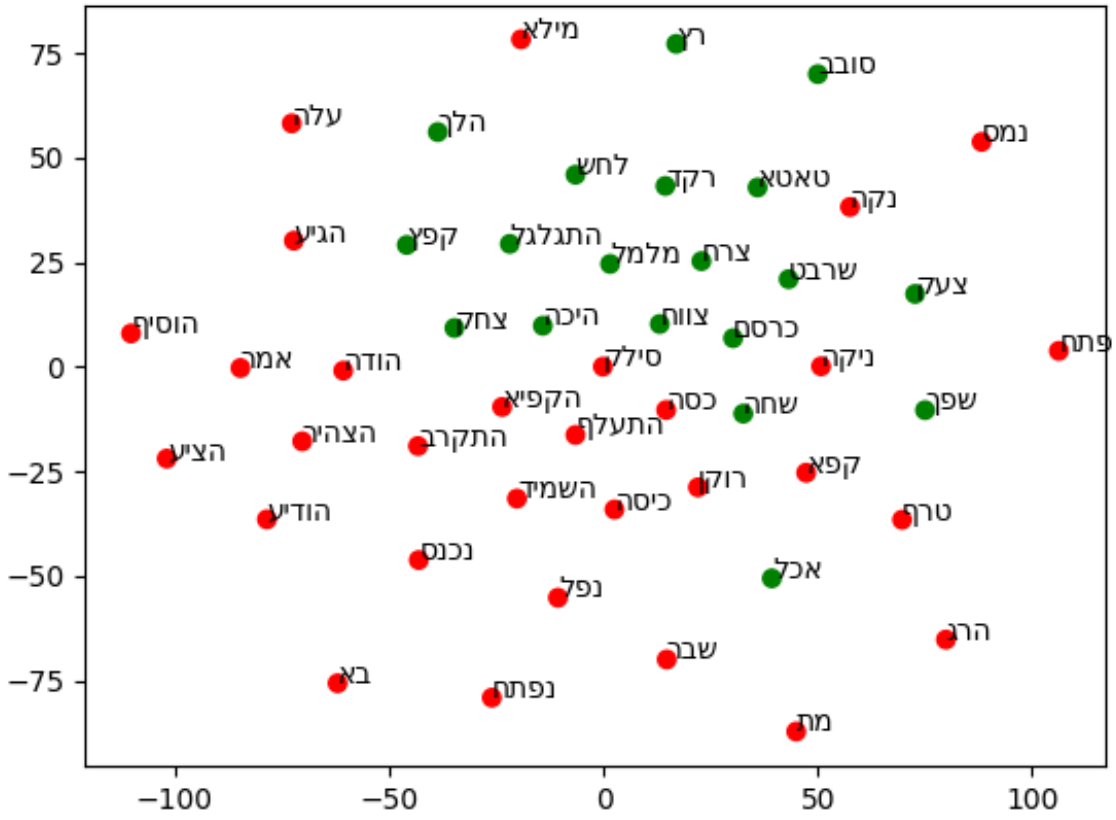
Experiment 2: Manner/Result in Hebrew



- Looks like a decent replication of the English experiment.
- Again semantic similarity seems to be reflected graphically.
- Also: some alternations are fairly close to one another (*open*).

Experiment 2: Manner/Result in Hebrew

- Ran `UDPipe` (Straka and Straková 2017) for the syntactic analysis.
- Using dependencies improves performance, as expected.
- The same caveats hold (no quantitative analysis, t-SNE is non-deterministic).



Experiment 2: Manner/Result in Hebrew

- Manner/Result has been argued to be a property of roots and not of verbs.
- Experiment 2b aimed to replicate the findings using Hebrew *roots* rather than verbs.

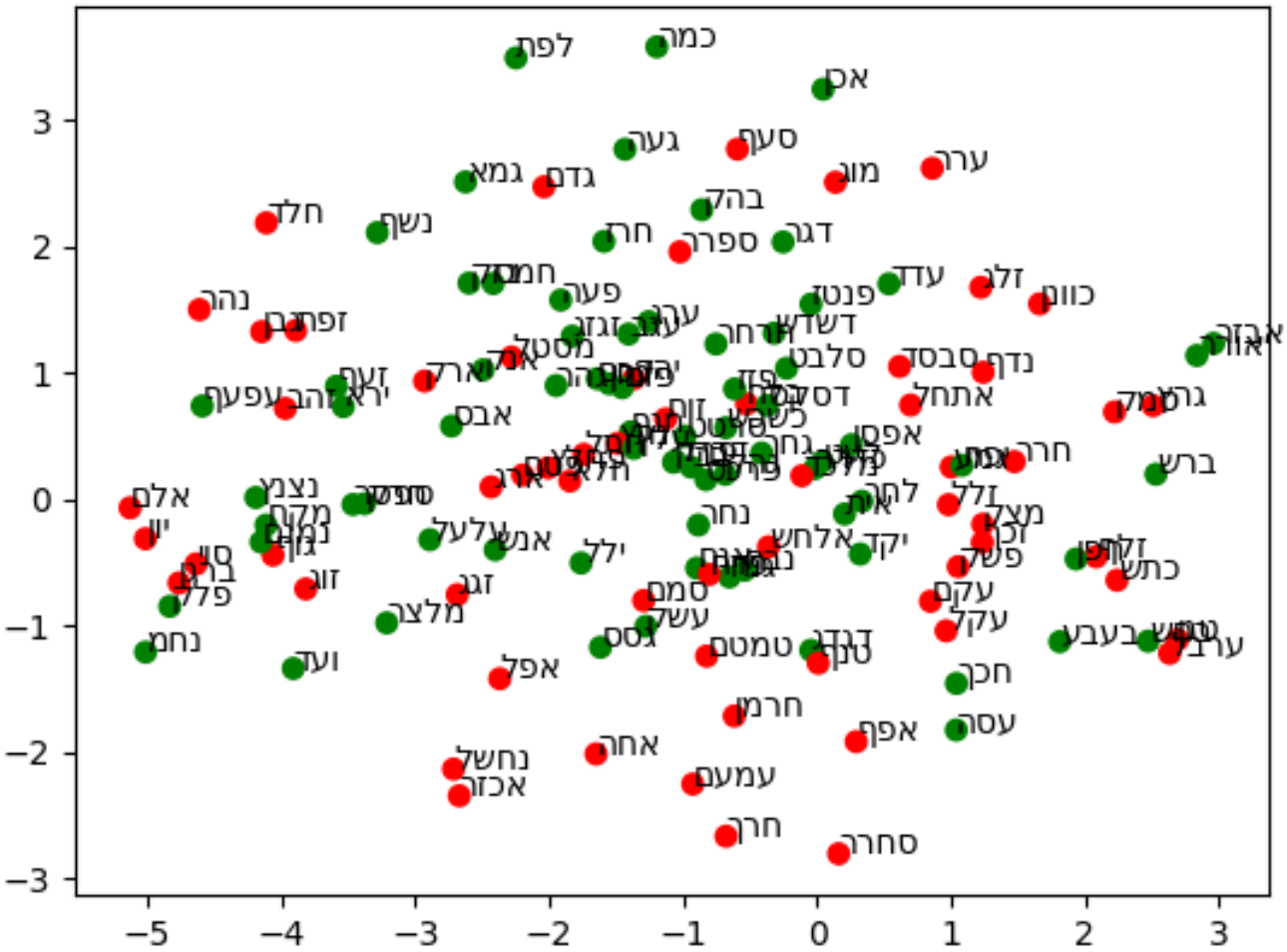
Experiment 2: Manner/Result in Hebrew

Experiment 2b:

1. Started with the Hebrew Wikipedia corpus (Itai and Wintner 2008).
2. Item selection for analysis:
   - Extracted all roots from the analyzed corpus.
   - Chose a random subset of 200 roots.
   - 70% high-frequency (>100 tokens), 30% low-frequency.
3. Coded roots as Manner/Result using diagnostics in the literature:
   - 48% Manner, 52% Result.
4. Parsed the corpus with `yap` (More et al. 2019).
   - Current state of the art for Hebrew parsing.
   - Not an RNN.
   - Produces a first approximation of roots and templates for each verb.
5. Let `word2vec` run on this corpus with roots instead of verbs.
   - *patax* 'opened', *niftax* 'got opened' $\rightarrow \sqrt{\text{PTX}}$.
   - *sagar* 'closed', *hesgir* 'extradited' $\rightarrow \sqrt{\text{SGR}}$.

Experiment 2: Manner/Result in Hebrew

- No clear separation when using roots rather than verbs.

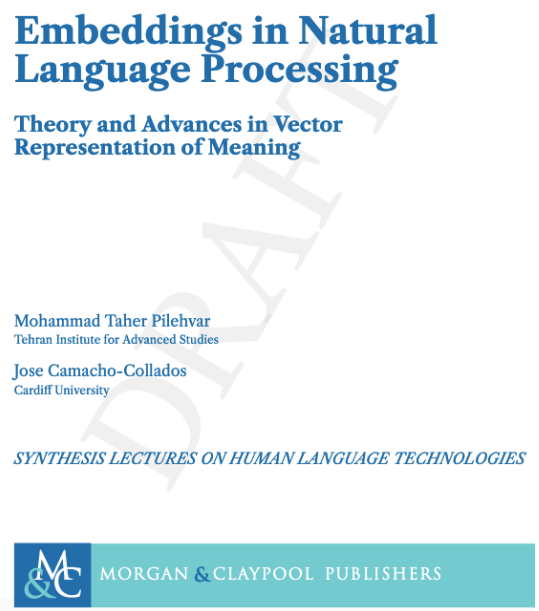


Experiment 2: Manner/Result in Hebrew

Qualitative summary of the qualitative analysis:

English	Words	✓
English	Dependencies	✓✓
Hebrew	Words	✓
Hebrew	Dependencies	✓✓
Hebrew	Roots	✗ ?

1. Kastner (2020a) explored "root embeddings":
   - Word embeddings are good psycholinguistic predictors in English and similar languages.
   - Terrible in Hebrew.
2. Still a lot of empirical work left to do in order to evaluate what these models are really learning, even for semantics.
3. It's something about similarity in distribution, rather than similarity in meaning.



Further reading

- Recent overviews: Lenci (2018); Boleda (2020).
- Potts and Petersen (2022): Lexical semantics in the time of large language models, `https://www.youtube.com/watch?v=EbwtZtd8XRo`
- Kastner (2020a): Farhy and Veríssimo (2019) used similarity ratings provided by Hebrew speakers to predict cross-modal priming. Word2vec couldn't do it.
- Utsumi (2020) and Grand et al. (2022): what lexical semantic properties are these models able to learn?
- Good work in distributional semantics which acknowledges the existence of lexical semantics *without* word2vec: Mitchell and Lapata (2010); Marelli and Baroni (2015); Vecchi et al. (2017); Varvara et al. (2021).
- And with word2vec: Mitchell and Steedman (2015); Pross et al. (2017).
- Ravfogel et al. (2020): learn embeddings, discard the lexical semantics, and keep the structural information, which means that they should be able to isolate the lexical semantics.
- `https://twitter.com/ryandcotterell/status/1556977691848491011`

1. Recap
2. Experimental: Are the patterns robust?
3. Experimental: Do they generalize?
4. Computational: Can they be derived from distributions?
5. Summary

Summary

1. What are the most robust crosslinguistic generalizations regarding the interaction between lexicon and grammar?
   - The root classes of Levinson (2007, 2010, 2014).
   - Different kinds of COS verbs.
2. (What formal tools can account for these?)
3. (Is it possible to reach a constrained inventory of lexical semantic primitives?)
4. How can these claims be tested experimentally and modeled computationally?
   - Comparative corpus work.
   - Acceptability studies.
   - Behavioral studies.
   - Models of distributional semantics – though distribution of what?

References I

Albright, Adam, and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90:119–161.

Alexiadou, Artemis, Fabienne Martin, and Florina Schäfer. 2017. Optionally causative manner verbs: when implied results get entailed. In *Roots V, UCL/QMUL*.

Anagnostopoulou, Elena. 2015. Exploring roots in their contexts: instrument verbs, manners and results in adjectival participles. In *Roots IV, New York University*.

Aronoff, Mark. 1994. *Morphology by itself: Stems and inflectional classes*. Cambridge, MA: MIT Press.

Beavers, John, Michael Everdell, Kyle Jerro, Henri Kauhanen, Andrew Koontz-Garboden, Elise LeBovidge, and Stephen Nichols. 2021. States and changes of state: A crosslinguistic study of the roots of verbal meaning. *Language* 97:439–484. URL `http://dx.doi.org/10.1353/lan.2021.0044`.

Berko, Jean. 1958. The child's learning of English morphology. *Word* 14:150–177.

Boleda, Gemma. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics* 6:213–234. URL `https://doi.org/10.1146%2Fannurev-linguistics-011619-030303`.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Clark, Eve V., and Herbert H. Clark. 1979. When nouns surface as verbs. *Language* 55:767–811.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/N19-1423`.

Dixon, R. M. W. 1982. *Where have all the adjectives gone?*. The Hague: Mouton.

Drummond, Alex. n.d. Ibex 0.3.8. Spellout.net/ibexfarm.

Ettinger, Allyson. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics* 8:34–48. URL `http://dx.doi.org/10.1162/tacl_a_00298`.

Ettinger, Allyson, and Tal Linzen. 2016. Evaluating vector space models using human semantic priming results. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 72–77. Berlin, Germany: Association for Computational Linguistics.

Farhy, Yael, and João Veríssimo. 2019. Semantic effects in morphological priming: The case of Hebrew stems. *Language and Speech* 62:737–750.

References II

Fisher, Cynthia, Yael Gertner, Rose M. Scott, and Sylvia Yuan. 2010. Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science* 1:143–149.

Gillette, Jane, Henry Gleitman, Lila R. Gleitman, and Anne Lederer. 1999. Human simulations of vocabulary learning. *Cognition* 73:135–176.

Gleitman, Lila R. 1990. The structural sources of verb meanings. *Language Acquisition* 1:3–55.

Grand, Gabriel, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour* 6:975–987. URL `http://dx.doi.org/10.1038/s41562-022-01316-8`.

Harley, Heidi. 2014. On the identity of roots. *Theoretical Linguistics* 40:225–276.

Hewitt, John, and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4129–4138. Minneapolis, Minnesota: Association for Computational Linguistics. URL `https://aclanthology.org/N19-1419`.

Heyman, Tom, Anke Bruninx, Keith A. Hutchison, and Gert Storms. 2018. The (un)reliability of item-level semantic priming effects. *Behavior Research Methods* 50:2173–2183. URL `http://dx.doi.org/10.3758/s13428-018-1040-9`.

Irwin, Patricia, and Itamar Kastner. 2020. Semantic primitives at the syntax-lexicon interface. Ms., Swarthmore College and University of Edinburgh. lingbuzz/005302.

Itai, Alon, and Shuly Wintner. 2008. Language resources for Hebrew. *Language Resources and Evaluation* 42:75–98.

Kastner, Itamar. 2020a. Predicting semantic priming in Hebrew morphology using word embeddings. In *AMLaP 2020*.

Kastner, Itamar. 2020b. *Voice at the interfaces: The syntax, semantics and morphology of the Hebrew verb*. Number 8 in Open Generative Syntax. Berlin: Language Science Press.

Kastner, Itamar, and Frans Adriaans. 2018. Linguistic constraints on statistical word segmentation: The role of consonants in Arabic and English. *Cognitive Science* 42:494–518.

Keller, Frank. 2010. Cognitively plausible models of human language processing. In *Proceedings of the ACL 2010 Conference Short Papers*, 60–67. Uppsala, Sweden: Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P10-2012`.

Kogkalidis, Konstantinos, and Gijs Wijnholds. 2022. Discontinuous constituency and BERT: A case study of Dutch. In *Findings of the association for computational linguistics*.

Kratzer, Angelika. 2000. Building statives. In *Proceedings of the twenty-sixth annual meeting of the Berkeley Linguistics Society*, ed. Lisa J. Conathan, Jeff Good, Darya Kavitskaya, Alyssa B. Wulf, and Alan C. L. Yu, 385–399. Berkeley, CA: University of California, Berkeley Linguistics Society.

References III

Landau, Barbara, and Lila R. Gleitman. 1985. *Language and experience: Evidence from the blind child*. Cambridge, MA: Harvard University Press.

Landauer, Thomas, and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104:211–240.

Lenci, Alessandro. 2018. Distributional models of word meaning. *Annual Review of Linguistics* 4:151–171.

Levin, Beth, and Malka Rappaport Hovav. 2005. *Argument realization*. Research Surveys in Linguistics Series. Cambridge, UK: Cambridge University Press.

Levinson, Lisa. 2007. The roots of verbs. Doctoral Dissertation, New York University, New York, NY.

Levinson, Lisa. 2010. Arguments for pseudo-resultative predicates. *Natural Language and Linguistic Theory* 28:135–182.

Levinson, Lisa. 2014. The ontology of roots and verbs. In *The syntax of roots and the roots of syntax*, ed. Artemis Alexiadou, Hagit Borer, and Florian Schäfer, 208–229. Oxford: Oxford University Press.

Levy, Omer, and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 302–308. Baltimore, Maryland.

Linzen, Tal. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 13–18. Berlin: Association for Computational Linguistics.

Linzen, Tal. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5210–5217. Online: Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.acl-main.465`.

Linzen, Tal, and Marco Baroni. 2020. Syntactic structure from deep learning. *Annual Review of Linguistics* .

van der Maaten, Laurens J. P., and Geoffrey E. Hinton. 2008. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research* 9:2579–2605.

Marelli, Marco, and Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review* 122:485–515.

Merkx, Marjolein, Kathleen Rastle, and Matthew H. Davis. 2011. The acquisition of morphological knowledge investigated through artificial language learning. *The Quarterly Journal of Experimental Psychology* 64:1200–1220.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781. URL `http://arxiv.org/abs/1301.3781`.

Mitchell, Jeff, and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science* 34:1388–1429. URL `http://dx.doi.org/10.1111/j.1551-6709.2010.01106.x`.

References IV

Mitchell, Jeff, and Mark Steedman. 2015. Orthogonality of syntax and semantics within distributional spaces. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1301–1310. Beijing, China: Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P15-1126`.

More, Amir, Amit Seker, Victoria Basmova, and Reut Tsarfaty. 2019. Joint transition-based models for morpho-syntactic parsing: Parsing strategies for MRLs and a case study from modern Hebrew. *Transactions of the Association for Computational Linguistics* 7:33–48. URL `https://www.aclweb.org/anthology/Q19-1003`.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. URL `http://www.aclweb.org/anthology/D14-1162`.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Procedings of NAACL*.

Pross, Tillmann, Antje Roßdeutscher, Sebastian Padó, Gabriella Lapesa, and Max Kisselew. 2017. Integrating lexical-conceptual and distributional semantics: a case report. In *Proceedings of the 21st Amsterdam Colloquium*, ed. A. Cremers, T. van Gessel, and Floris Roelofsen, 75–85.

Rappaport Hovav, Malka. 2017. Grammatically relevant ontological categories underlie manner/result complementarity. In *Proceedings of IATL 32*, ed. Noa Brandel, volume 86, 77–98. MITWPL.

Rappaport Hovav, Malka, and Beth Levin. 2010. Reflections on manner/result complementarity. In *Syntax, lexical semantics, and event structure*, ed. Edit Doron, Malka Rappaport Hovav, and Ivy Sichel, 21–38. Oxford: Oxford University Press.

Ravfogel, Shauli, Yanai Elazar, Jacob Goldberger, and Yoav Goldberg. 2020. Unsupervised distillation of syntactic information from contextualized word representations. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 91–106. Association for Computational Linguistics. URL `https://aclanthology.org/2020.blackboxnlp-1.9`.

Roßdeutscher, Antje, and Hans Kamp. 2010. Syntactic and semantic constraints in the formation and interpretation of *ung*-nouns. In *Nominalisations across languages and frameworks*, ed. Artemis Alexiadou and Monika Rathert. Berlin: Mouton de Gruyter.

Spathas, Giorgos, and Dimitris Michelioudakis. 2020. States in the decomposition of verbal predicates. *Natural Language & Linguistic Theory* 39:1253–1306. URL `http://dx.doi.org/10.1007/s11049-020-09496-6`.

Straka, Milan, and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99. Vancouver, Canada: Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/K/K17/K17-3009.pdf`.

Tamminen, J., M. H. Davis, and K Rastle. 2015. From specific examples to general knowledge in language learning. *Cognitive Psychology* 79:1–39.

References V

Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*. URL `https://openreview.net/forum?id=SJzSgnRcKX`.

Utsumi, Akira. 2020. Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science* 44. URL `http://dx.doi.org/10.1111/cogs.12844`.

Varvara, Rossella, Gabriella Lapesa, and Sebastian Padó. 2021. Grounding semantic transparency in context: A distributional semantic study on German event nominalizations. *Morphology* 31:409–446. URL `http://dx.doi.org/10.1007/s11525-021-09382-w`.

Vecchi, Eva Maria, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2017. Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive Science* 41:102–136.

Řehůřek, Radim, and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

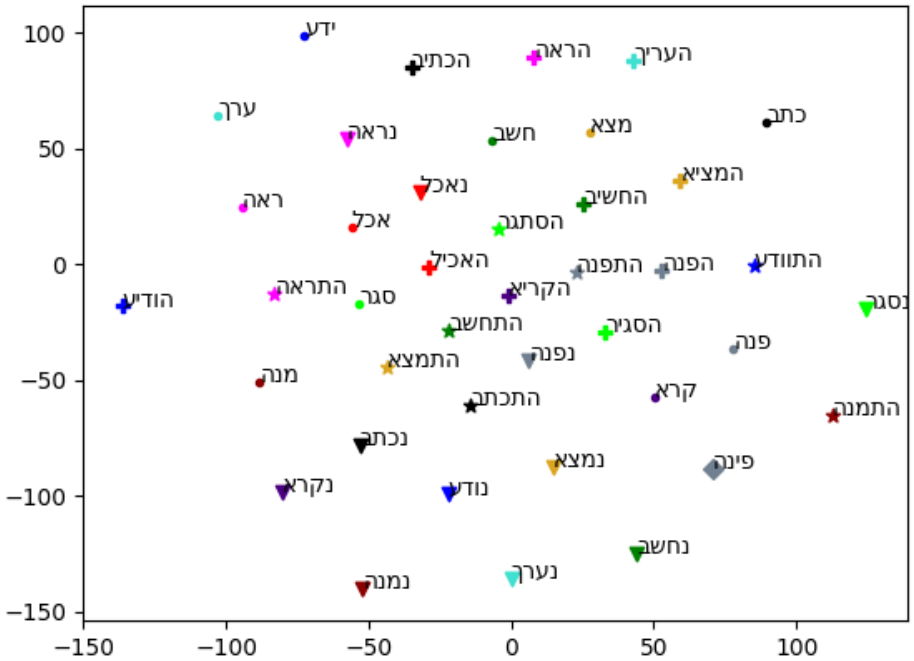
Distance in alternations within a root

1. How much can the meaning of a root vary across alternations? (Harley 2014)

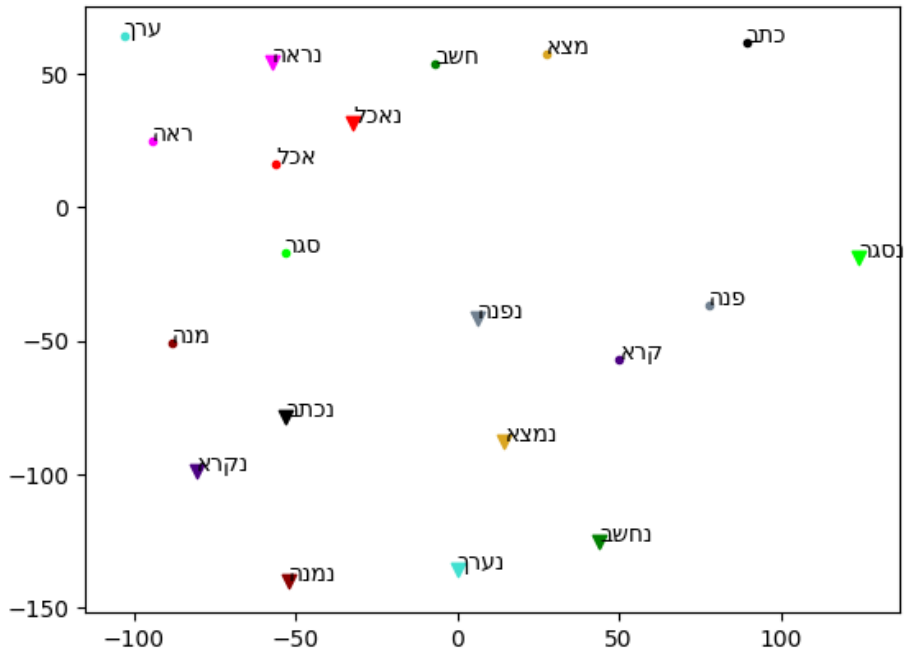
	niXYaZ		*XaYaZ*		*heXYiZ*	
$\sqrt{\text{KTB}}$	*nixtav*	'was written'	*katav*	'wrote'	*hextiv*	'dictated'
$\sqrt{\text{KRJ}}$	*nikra*	'was read'	*kara*	'read'	*hekri*	'read out'
$\sqrt{\text{SGR}}$	*nisgar*	'was closed'	*sagar*	'closed'	*hesgir*	'extradited'

2. Tested a subset of roots and templates.

Subset of roots



Subset of roots, only *XaYaZ*∼*niXYaZ*



Experiment 3: Background

Hebrew

- Complex non-concatenative morphology.
- Relationship between word forms and between lemmas is not derived by simple affixation.

Farhy and Veríssimo (2019): cross-modal priming.

Form	Prime			Target
	Unrelated	*XaYaZ*	*XiYeZ*	*hitXaYeZ*
Infinitive	*lʃpr* /leʃaper/ 'improve'	*lxlwq* /laxlok/ 'share s.th.'	*lxlq* /lexalek/ 'divide s.th.'	*htxlq* /hitxalek/ 'got divided'
1st Past	*bjʃltj* /biʃalti/ 'cooked'	*lvvjtj* /laviti/ 'borrowed'	*ljvvjtj* /liviti/ 'accompanied'	*htlvvh* /hitlava/ 'joined'

- **Obtained** human ratings of relatedness between primes and targets.
- **Finding**: interaction of Template and Relatedness.
- **Today**: attempt to replicate using embeddings. (Ettinger and Linzen 2016)

Experiment 3: Background

Hebrew

- Complex non-concatenative morphology.
- Relationship between word forms and between lemmas is not derived by simple affixation.

Farhy and Veríssimo (2019): cross-modal priming.

Form	Prime			Target
	Unrelated	*XaYaZ*	*XiYeZ*	*hitXaYeZ*
Infinitive	*lʃpr*	*lxlwq*	*lxlq*	*htxlq*
	/leʃaper/	/laxlok/	/lexalek/	/hitxalek/
	‘improve’	‘share s.th.’	‘divide s.th.’	‘got divided’
1st Past	*bjʃltj*	*lvvjtj*	*ljvvjtj*	*htlvvh*
	/biʃalti/	/laviti/	/liviti/	/hitlava/
	‘cooked’	‘borrowed’	‘accompanied’	‘joined’

- **Obtained** human ratings of relatedness between primes and targets.
- **Finding**: interaction of Template and Relatedness.
- **Today**: attempt to replicate using embeddings. (Ettinger and Linzen 2016)

Experiment 3: Background

Hebrew

- Complex non-concatenative morphology.
- Relationship between word forms and between lemmas is not derived by simple affixation.

Farhy and Veríssimo (2019): cross-modal priming.

Form	Prime			Target
	Unrelated	*XaYaZ*	*XiYeZ*	*hitXaYeZ*
Infinitive	*lʃpr* /leʃaper/ 'improve'	*lxlwq* /laxlok/ 'share s.th.'	*lxlq* /lexalek/ 'divide s.th.'	*htxlq* /hitxalek/ 'got divided'
1st Past	*bjʃltj* /biʃalti/ 'cooked'	*lvvjtj* /laviti/ 'borrowed'	*ljvvjtj* /liviti/ 'accompanied'	*htlvvh* /hitlava/ 'joined'

- **Obtained** human ratings of relatedness between primes and targets.
- **Finding:** interaction of Template and Relatedness.
- **Today**: attempt to replicate using embeddings. (Ettinger and Linzen 2016)

Experiment 3: Background

Hebrew

- Complex non-concatenative morphology.
- Relationship between word forms and between lemmas is not derived by simple affixation.

Farhy and Veríssimo (2019): cross-modal priming.

Form	Prime			Target
	Unrelated	*XaYaZ*	*XiYeZ*	*hitXaYeZ*
Infinitive	*lʃpr* /leʃaper/ 'improve'	*lxlwq* /laxlok/ 'share s.th.'	*lxlq* /lexalek/ 'divide s.th.'	*htxlq* /hitxalek/ 'got divided'
1st Past	*bjʃltj* /biʃalti/ 'cooked'	*lvvjtj* /laviti/ 'borrowed'	*ljvvjtj* /liviti/ 'accompanied'	*htlvvh* /hitlava/ 'joined'

- **Obtained** human ratings of relatedness between primes and targets.
- **Finding**: interaction of Template and Relatedness.
- **Today:** attempt to replicate using embeddings. (Ettinger and Linzen 2016)

Experiment 3: Methods

1. **Regressed the original raw results** against models' similarity ratings.
2. **Embedding models** (with associated datasets):
   - w2vWords: simple word2vec. (Mikolov et al. 2013)
   - UDPipe syntactic dependencies with word2vec. (Straka and Straková 2017)
   - yap syntactic dependencies with word2vec. (More et al. 2019)
   - bert: ratings from multilingual BERT. (Devlin et al. 2019)
3. Model parameters: similar to earlier work on English.
   - Skip-gram.
   - 200 dimensions.
   - Window size 5.
4. Training data:
   - w2v models trained on the raw Hebrew data for the CoNLL 2017 Shared Task (615M words).
   - Multilingual BERT is pre-trained and was used as-is.
5. **Similarity ratings** between prime and target were calculated using cosine (and Pearson's) correlation.
6. **Regression**: RT $\sim$ Relatedness + Template + Relatedness:Template + (1 | Participant) + (1 | Item)

Experiment 3: Methods

1. **Regressed the original raw results** against models' similarity ratings.
2. **Embedding models** (with associated datasets):
   - w2vWords: simple word2vec. (Mikolov et al. 2013)
   - UDPipe syntactic dependencies with word2vec. (Straka and Straková 2017)
   - yap syntactic dependencies with word2vec. (More et al. 2019)
   - bert: ratings from multilingual BERT. (Devlin et al. 2019)
3. Model parameters: similar to earlier work on English.
   - Skip-gram.
   - 200 dimensions.
   - Window size 5.
4. Training data:
   - w2v models trained on the raw Hebrew data for the CoNLL 2017 Shared Task (615M words).
   - Multilingual BERT is pre-trained and was used as-is.
5. **Similarity ratings** between prime and target were calculated using cosine (and Pearson's) correlation.
6. **Regression**: RT $\sim$ Relatedness + Template + Relatedness:Template + (1 | Participant) + (1 | Item)

Experiment 3: Methods

1. **Regressed the original raw results** against models' similarity ratings.
2. **Embedding models** (with associated datasets):
   - w2vWords: simple word2vec. (Mikolov et al. 2013)
   - UDPipe syntactic dependencies with word2vec. (Straka and Straková 2017)
   - yap syntactic dependencies with word2vec. (More et al. 2019)
   - BERT: ratings from multilingual BERT. (Devlin et al. 2019)
3. Model parameters: similar to earlier work on English.
   - Skip-gram.
   - 200 dimensions.
   - Window size 5.
4. Training data:
   - w2v models trained on the raw Hebrew data for the CoNLL 2017 Shared Task (615M words).
   - Multilingual BERT is pre-trained and was used as-is.
5. **Similarity ratings** between prime and target were calculated using cosine (and Pearson's) correlation.
6. **Regression**: RT $\sim$ Relatedness + Template + Relatedness:Template + (1 | Participant) + (1 | Item)

Experiment 3: Methods

1. **Regressed the original raw results** against models' similarity ratings.
2. **Embedding models** (with associated datasets):
   - w2vWords: simple word2vec. (Mikolov et al. 2013)
   - UDPipe syntactic dependencies with word2vec. (Straka and Straková 2017)
   - yap syntactic dependencies with word2vec. (More et al. 2019)
   - BERT: ratings from multilingual BERT. (Devlin et al. 2019)
3. Model parameters: similar to earlier work on English.
   - Skip-gram.
   - 200 dimensions.
   - Window size 5.
4. Training data:
   - w2v models trained on the raw Hebrew data for the CoNLL 2017 Shared Task (615M words).
   - Multilingual BERT is pre-trained and was used as-is.
5. **Similarity ratings** between prime and target were calculated using cosine (and Pearson's) correlation.
6. **Regression:** RT $\sim$ Relatedness + Template + Relatedness:Template + (1 | Participant) + (1 | Item)

Experiment 3

- Each model performs its own lemmatization.
- So the models ended up with different vocabularies.
- The original experimental dataset (FV19) had out-of-vocabulary items.
- To address this, each model was evaluated on two datasets:
  1. Its "own" dataset: intersection of the model's vocabulary and FV19.
  2. Smaller dataset containing only the items shared by all four models.
- BERT had no out-of-vocabulary items (BERT dataset = FV19).

Dataset	Primes	Targets
FV19	82	41
w2vWords	50	29
UDPipe	31	20
yap	39	24
no-OOV	30	20

Experiment 3: Results

First of all, Cosine correlations between ratings:

	Human	w2vWords	UDPipe	yap
w2vWords	0.0855	—	—	—
UDPipe	−0.0145	0.247	—	—
yap	−0.0597	0.365	**0.7916**	—
bert	0.0306	0.2084	−0.0002	0.1557

- Nothing correlates even remotely well with human ratings.
- This might temper your enthusiasm for the results.

Experiment 3: Results

First of all, Cosine correlations between ratings:

	Human	w2vWords	UDPipe	yap
w2vWords	0.0855	—	—	—
UDPipe	−0.0145	0.247	—	—
yap	−0.0597	0.365	**0.7916**	—
bert	0.0306	0.2084	−0.0002	0.1557

- Nothing correlates even remotely well with human ratings.
- This might temper your enthusiasm for the results.

Experiment 3: Results

- Farhy and Veríssimo (2019):
  1. Effect of **Semantic Relatedness** between prime and target.
  2. **Interaction** of Semantic Relatedness and Prime Type (verbal template).
- The crucial metric is the t-value of the interaction ($t > 2 \approx$ significant).

	Dataset	HUMAN	W2V	UDPIPE	YAP	BERT
SemRel	FV19	**2.469**	—	—	—	0.527
	ds-w2vWords	**2.807**	0.834	—	—	0.437
	ds-UDPipe	0.748	—	0.102	—	1.405
	ds-yap	1.472	—	—	0.053	0.461
	ds-no-OOV	1.548	0.474	0.624	0.037†	0.853
SemRel:Tmplt	FV19	**2.497**	—	—	—	1.081
	ds-w2vWords	**3.059**	0.045	—	—	0.088
	ds-UDPipe	**2.766**	—	0.207	—	0.34
	ds-yap	**3.08**	—	—	0.398	0.703
	ds-no-OOV	**3.12**	0.151	0.318	0.541†	0.134

$\Rightarrow$ Human ratings show the finding; no replication with word embeddings.

Experiment 3: Results

- Farhy and Veríssimo (2019):
  1. Effect of **Semantic Relatedness** between prime and target.
  2. **Interaction** of Semantic Relatedness and Prime Type (verbal template).
- The crucial metric is the t-value of the interaction ($t > 2 \approx$ significant).

	Dataset	HUMAN	W2V	UDPIPE	YAP	BERT
SemRel	FV19	**2.469**	—	—	—	0.527
	ds-w2vWords	**2.807**	0.834	—	—	0.437
	ds-UDPipe	0.748	—	0.102	—	1.405
	ds-yap	1.472	—	—	0.053	0.461
	ds-no-OOV	1.548	0.474	0.624	0.037†	0.853
SemRel:Tmplt	FV19	**2.497**	—	—	—	1.081
	ds-w2vWords	**3.059**	0.045	—	—	0.088
	ds-UDPipe	**2.766**	—	0.207	—	0.34
	ds-yap	**3.08**	—	—	0.398	0.703
	ds-no-OOV	**3.12**	0.151	0.318	0.541†	0.134

⇒ Human ratings show the finding; no replication with word embeddings.

Experiment 3: Discussion

Summary

- Original experiment: an observation based on linguistic study fed directly into an experimental prediction.
- Original finding was based on human ratings.
- Reanalysis using word embeddings: null result.

1. Additional parameters?
   - Worth trying! But why don't the English models need additional firepower? What about Hebrew-learning children?
2. Different segmentation? Something like fastText might work.
   - Worth trying! But why don't the English models need this?
   - Not sure even that would be enough (Kastner and Adriaans 2018).
3. Wrong evaluation technique?
   - It's unclear to what extent human similarity ratings predict priming latencies (Heyman et al. 2018), or should predict cross-modal priming results.
4. Maybe word embeddings are not about meaning.

   ⇒ Or maybe we need to move away from traditional evaluation paradigms and towards human-like generalization. (Keller 2010; Linzen 2020)

Experiment 3: Discussion

Summary

- Reanalysis using word embeddings: null result.

1. Additional parameters?
   - Worth trying! But why don't the English models need additional firepower? What about Hebrew-learning children?
2. Different segmentation? Something like fastText might work.
   - Worth trying! But why don't the English models need this?
   - Not sure even that would be enough (Kastner and Adriaans 2018).
3. Wrong evaluation technique?
   - It's unclear to what extent human similarity ratings predict priming latencies (Heyman et al. 2018), or should predict cross-modal priming results.
4. Maybe word embeddings are not about meaning.

   ⇒ Or maybe we need to move away from traditional evaluation paradigms and towards human-like generalization. (Keller 2010; Linzen 2020)

Experiment 3: Discussion

Summary

- Reanalysis using word embeddings: null result.

1. Additional parameters?
   - Worth trying! But why don't the English models need additional firepower? What about Hebrew-learning children?
2. Different segmentation? Something like fastText might work.
   - Worth trying! But why don't the English models need this?
   - Not sure even that would be enough (Kastner and Adriaans 2018).
3. Wrong evaluation technique?
   - It's unclear to what extent human similarity ratings predict priming latencies (Heyman et al. 2018), or should predict cross-modal priming results.
4. Maybe word embeddings are not about meaning.

   ⇒ Or maybe we need to move away from traditional evaluation paradigms and towards human-like generalization. (Keller 2010; Linzen 2020)

Experiment 3: Discussion

Summary

- Reanalysis using word embeddings: null result.

1. Additional parameters?
   - Worth trying! But why don't the English models need additional firepower? What about Hebrew-learning children?
2. Different segmentation? Something like fastText might work.
   - Worth trying! But why don't the English models need this?
   - Not sure even that would be enough (Kastner and Adriaans 2018).
3. Wrong evaluation technique?
   - It's unclear to what extent human similarity ratings predict priming latencies (Heyman et al. 2018), or should predict cross-modal priming results.
4. Maybe word embeddings are not about meaning.

   ⇒ Or maybe we need to move away from traditional evaluation paradigms and towards human-like generalization. (Keller 2010; Linzen 2020)

Experiment 3: Discussion

Summary

- Reanalysis using word embeddings: null result.

1. Additional parameters?
   - Worth trying! But why don't the English models need additional firepower? What about Hebrew-learning children?
2. Different segmentation? Something like fastText might work.
   - Worth trying! But why don't the English models need this?
   - Not sure even that would be enough (Kastner and Adriaans 2018).
3. Wrong evaluation technique?
   - It's unclear to what extent human similarity ratings predict priming latencies (Heyman et al. 2018), or should predict cross-modal priming results.
4. Maybe word embeddings are not about meaning.

   $\Rightarrow$ Or maybe we need to move away from traditional evaluation paradigms and towards human-like generalization. (Keller 2010; Linzen 2020)