What's in a verb class? Experimental and computational approaches

Itamar Kastner

University of Edinburgh

LEL Seminar Series
University of Manchester
🚩1 May 2025 🚩

Joint work with Paolo Cassina (ENS), Dan Lassiter, Rob Truswell

What's a verb class?

- At what level is it specified?
- What are the relevant properties?
  - Syntactic
  - Semantic
  - World knowledge
  - Frequency
  - ...
- What's consistent crosslinguistically?
- What are the structural primitives (morphemes/features/functions/operators)?

Typical contrast between *eat* (Manner) and *devour* (Result):

Object drop:

a. Sam ate and ate. # Sam devoured and devoured.

Non-agentive subject:

b. # The erosion ate the coastline. The erosion devoured the coastline.

Today

1. Focus on Manner/Result.
2. Investigate some diagnostics more thoroughly.
3. See what we can learn from computational models.
4. Extend to other verb classes.

Testing Manner/Result tests

Typical contrast between *eat* (Manner) and *devour* (Result):

(Levin and Rappaport Hovav 1991, 2005, 2013; Rappaport Hovav and Levin 1998, 2010; Beavers and Koontz-Garboden 2012, 2017, 2020; Mateu and Acedo-Matellán 2012; Acedo-Matellán and Mateu 2014; Rappaport Hovav 2017; Melchin 2019; Ausensi 2023), ...

Object drop:

a. Sam ate and ate. # Sam devoured and devoured.

Partial completion:

b. Sam ate the apple halfway. # Sam devoured the apple halfway.

Non-agentive subject:

c. # The erosion ate the coastline. The erosion devoured the coastline.

Out-prefixation:

d. ? Sam out-ate the other contestants. ?? Sam out-devoured the other contestants.

$\Rightarrow$ Manner/Result is distinguished by some linguistic diagnostics.

Testing Manner/Result tests

Manner, Result, and...

(Beavers and Koontz-Garboden 2017, 2020; Ausensi 2023)

- Verbs of cooking?
- Verbs of directed throwing?
- Verbs of killing?
- Verbs of stealing?

(1)
a. #Jesse braised and braised. ⇒ Result!
b. ?Jesse braised the cabbage halfway. ⇒ ???
c. #The pot braised the chard. ⇒ Manner!
d. Jesse out-braised the other chef. ⇒ Manner!

1. Is Manner/Result the ontology itself? Is the ontology about scales?
2. Are Cooking/Throwing/Killing/Stealing defined at the same level?
3. What about any other verb class?
4. How can we tell what class a given verb is in?

Testing Manner/Result tests

1. Is Manner/Result the ontology itself? Is the ontology about scales?
2. Are Cooking/Throwing/Killing/Stealing defined at the same level?
3. What about any other verb class?
4. How can we tell what class a given verb is in?

Our pilot study

- Tested six Manner/Result diagnostics in an acceptability study.
- Resultatives and denied change are the most robust.
- Perhaps the first study that lets us evaluate syntactic, semantic, pragmatic and lexical aspects of standard diagnostics.

Testing Manner/Result tests

- Manner/Result Complementarity: the claim that a given verb (or perhaps root) lexicalizes the manner or result of an action, but not both.
- Different researchers rely on different diagnostics, implemented in different ways.
- Difficult to apply the same set of considerations when extending the investigation to additional verb classes.
- Hard to tell what a given diagnostic is ultimately targeting.

Object drop

Manner verbs can drop their objects, but:

- Clauses consisting of only subject and predicate can sound unnatural.
- Some published examples include an adverbial phrase to help:

(2) The backpackers climbed all day. (Levin and Rappaport Hovav 2013:(25b))

- Others don't:

(3) #The toddler broke. (Rappaport Hovav and Levin 2010:(3a))

Testing Manner/Result tests

Object Drop

(Levin and Rappaport Hovav 2013:(25b))

(4) The backpackers climbed all day.

No Change

(Rappaport Hovav and Levin 2010:(4c))

(5) Chris scrubbed the tub for hours, but it didn't get any cleaner.

No Action

(Beavers and Koontz-Garboden 2012:(47))

(6) ?Isaac tossed the kids the balls after 4pm but didn't move a muscle.
Rather, he failed to stop the ball machine at the specified time.

Out-prefixation

(Beavers and Koontz-Garboden 2012:(19b))

(7) ?Kim outshattered the other bottle-shatterer.

Resultatives

(Rappaport Hovav and Levin 2010:(2b))

(8) Cinderella scrubbed her fingers raw.

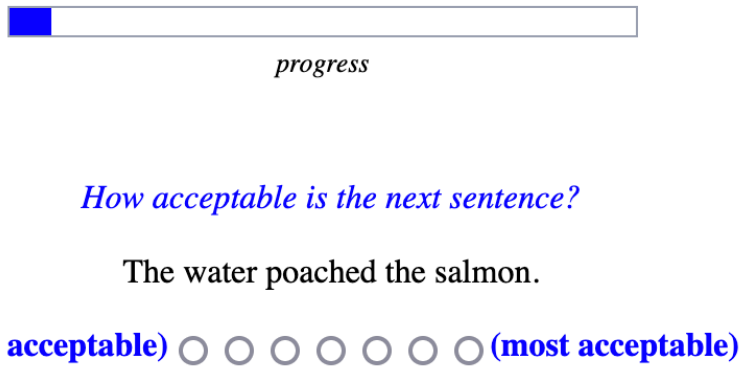
Subject/object ("selectional restrictions")

(Beavers and Koontz-Garboden 2012:(32c))

(9) #The earthquake wiped the floor.

Testing Manner/Result tests: Methods

1. Online acceptability study on PCIbex. (Zehr and Schwarz 2018; Drummond n.d)
2. Items presented on 7-point likert scale, labelled at the edges.



3. Three practice trials before the main experiment.
4. Grammatical and ungrammatical controls and fillers.
5. Order of all items randomized.
6. 48 participants (46 after exclusions).

Testing Manner/Result tests: Methods

- Three Manner verbs: *scrub, slam, wipe*
- Three Result verbs: *break, cut, shatter*
- Two Cooking verbs: *braise, poach*
- Two Throwing verbs: *throw, toss*
- Four Other verbs for comparison: *bang, know, sleep, yell*

Testing Manner/Result tests: Results by diagnostic

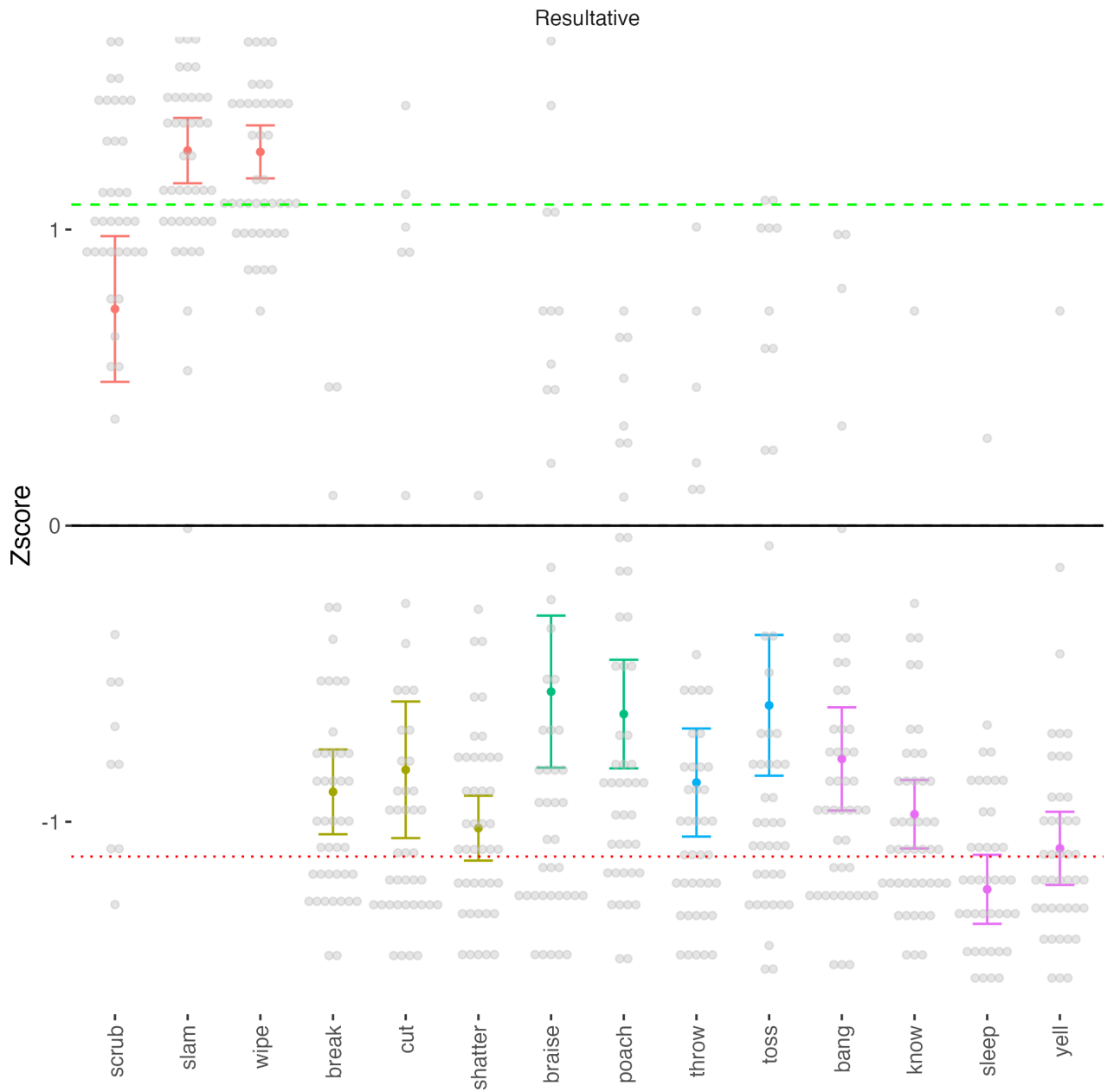
M Chris scrubbed her fingers raw.

R Kim broke her hands bloody.

C Jessie braised the chard burnt.

T Angus threw the tin dented.

O Ray banged the drum torn.



Testing Manner/Result tests: Results by diagnostic

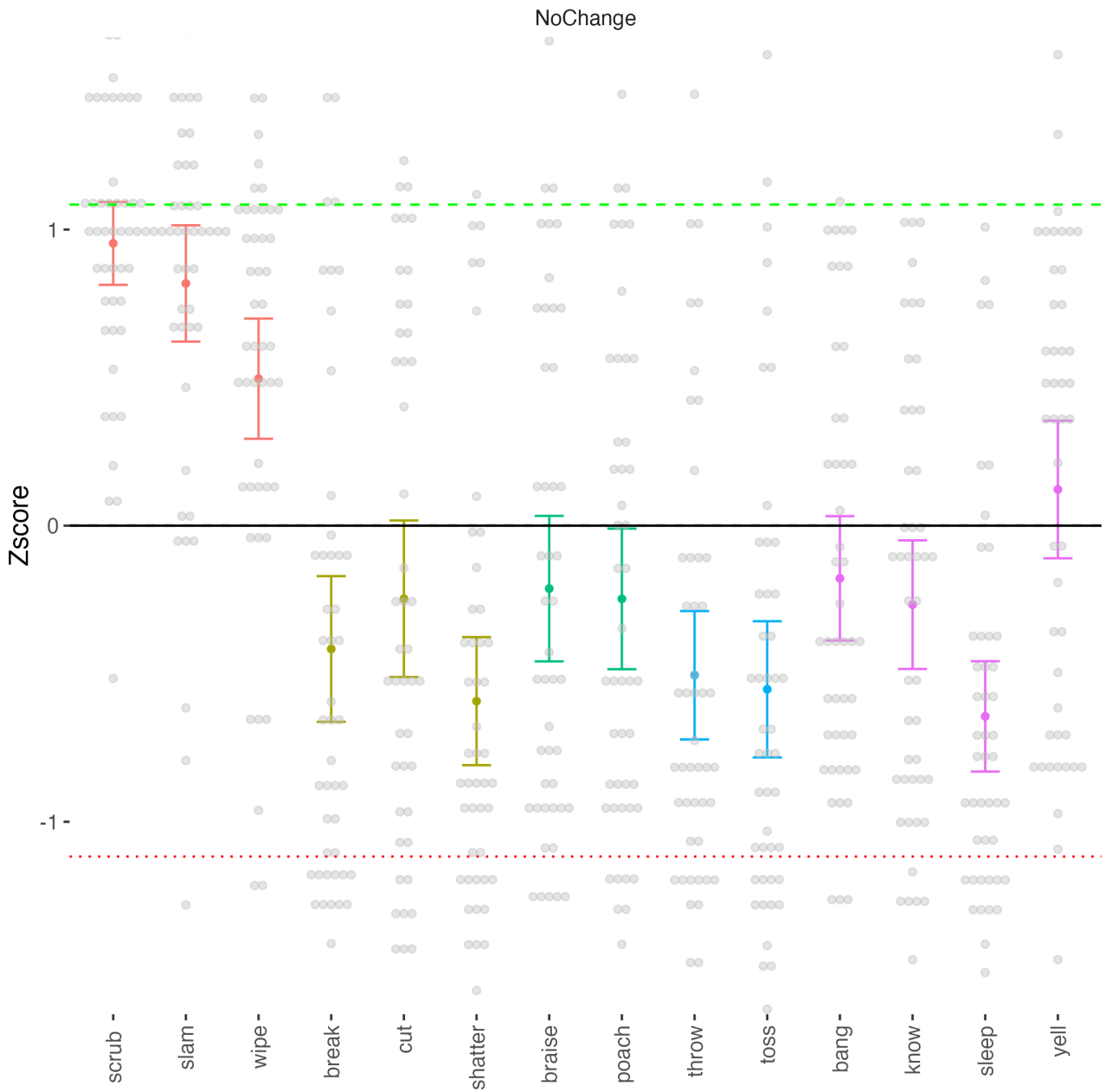
M Chris just scrubbed the tub, but it didn't get any cleaner.

R Kim just broke the vase, but nothing is different about it.

C Jessie just braised the chard, but nothing is different about it.

T Angus just threw Riley the tin, but it is not in a different place.

O Ray just banged the drum, but nothing is different about it.

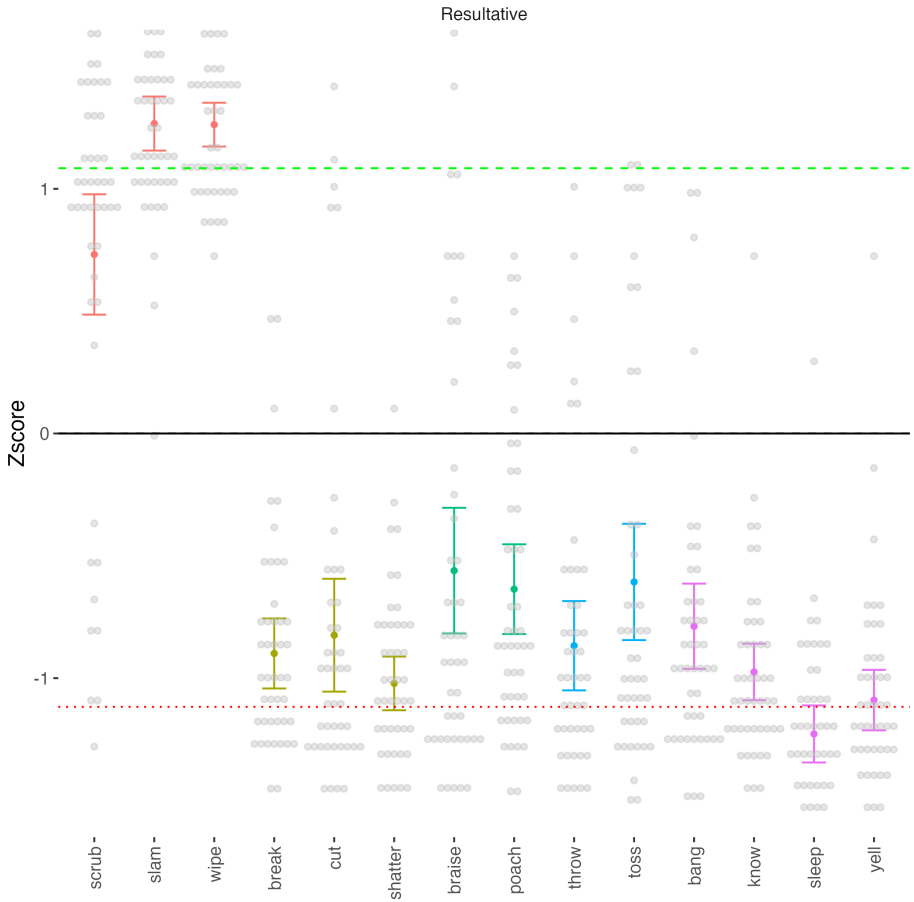


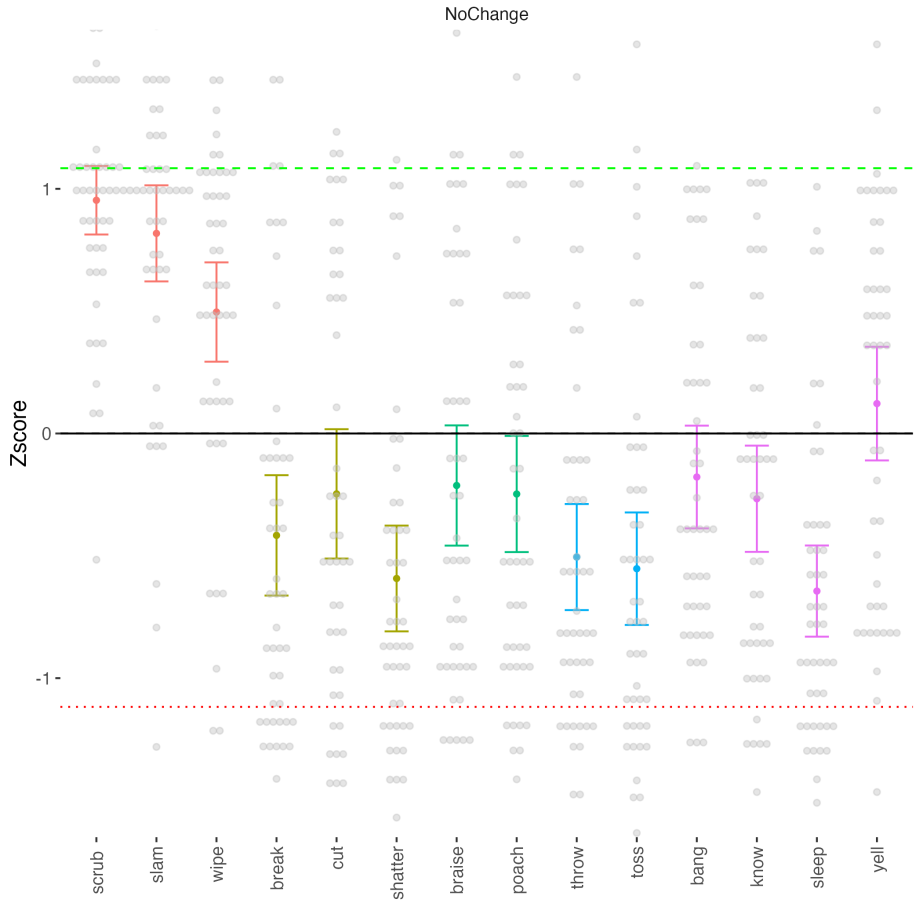
Testing Manner/Result tests: Results by diagnostic

Resultatives and No Change:

➤ Good discriminators of Manner/Result.

➤ Cooking and Throwing pattern with Result.





Testing Manner/Result tests: Results by diagnostic

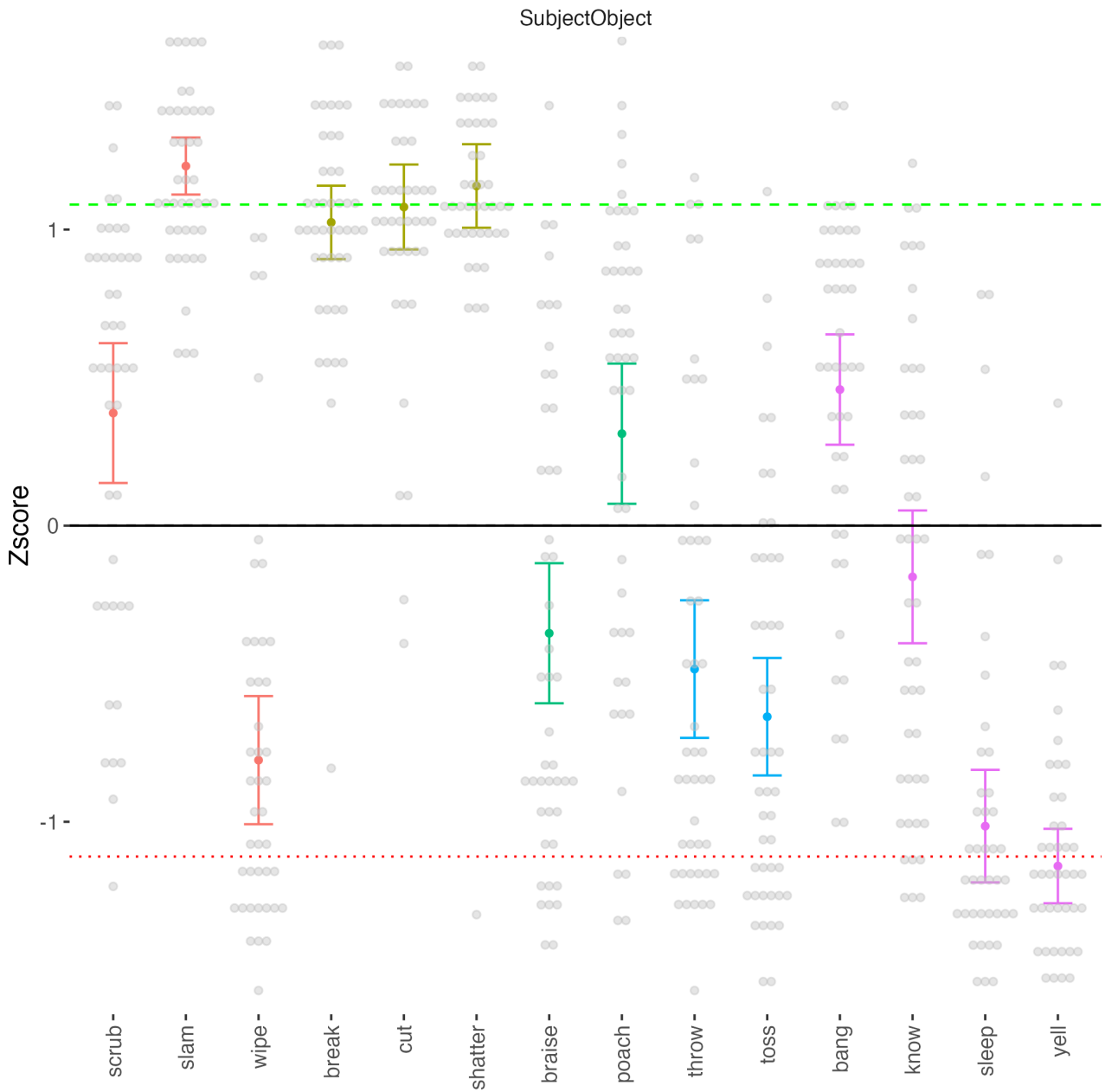
M The stiff brush scrubbed the tub.

R The hammer broke the vase.

C The heatwave braised the chard.

T The momentum threw the tin.

O The stick banged the drum.



Testing Manner/Result tests: Results by diagnostic

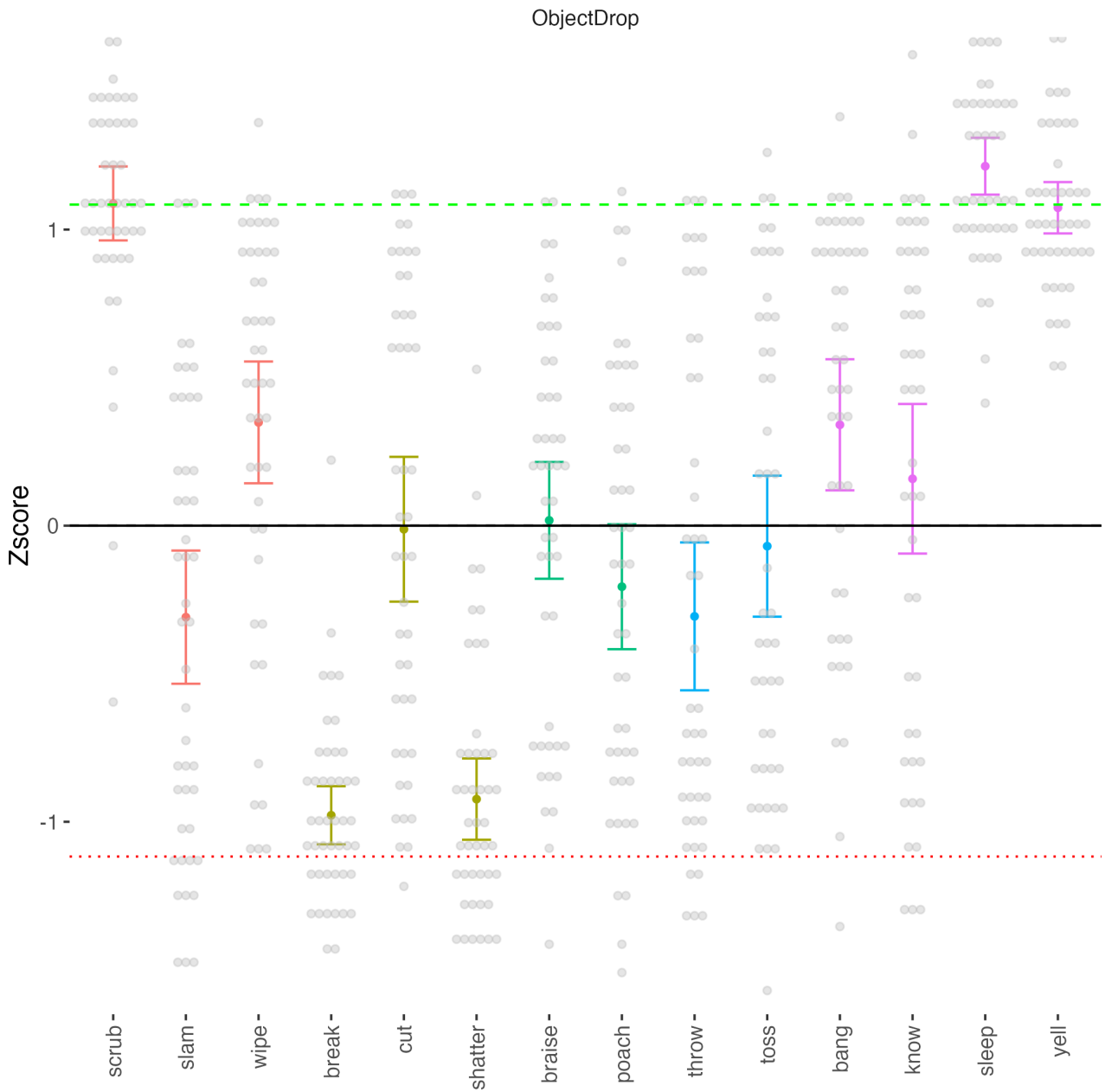
M Chris scrubbed all morning long.

R Kim broke all morning long.

C Jessie braised all morning long.

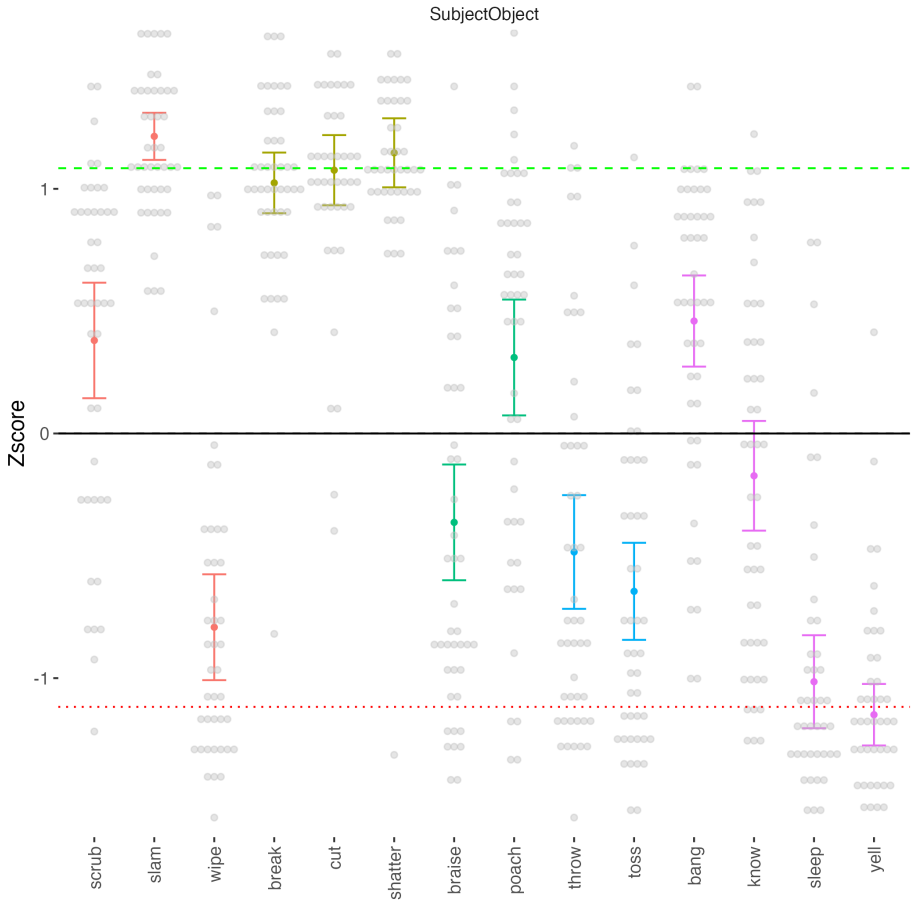
T Angus threw all morning long.

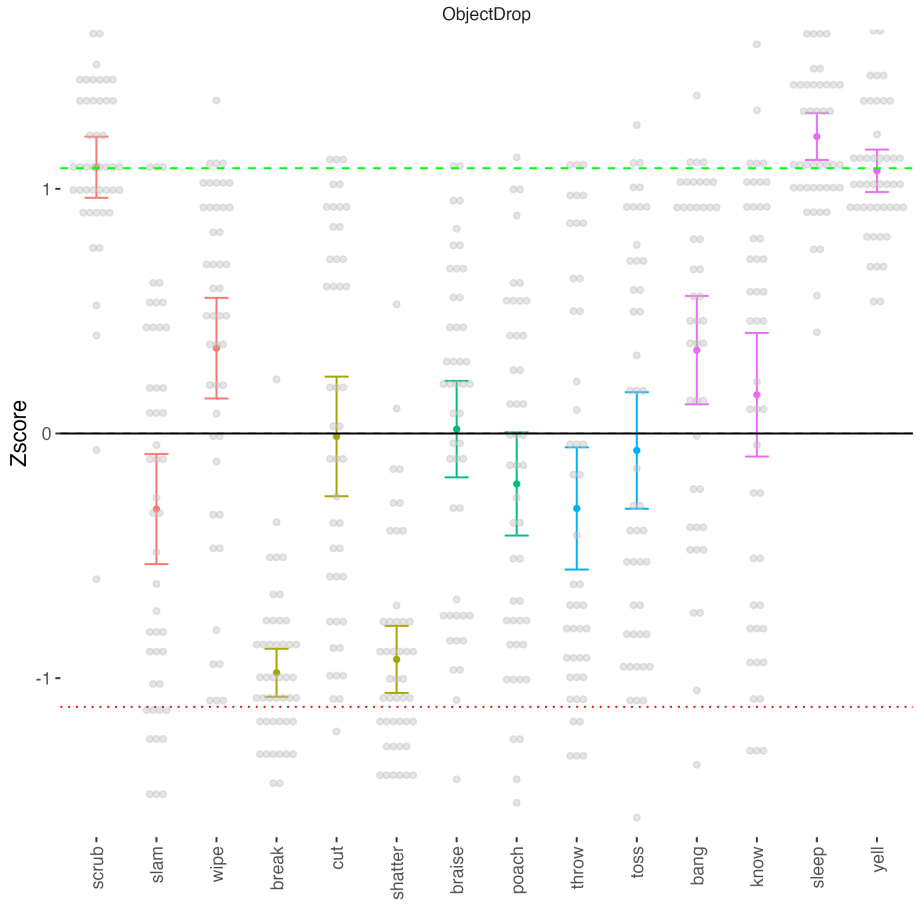
O Ray banged all morning long.



Testing Manner/Result tests: Results by diagnostic

Subject sensitivity (agenthood) and Object Drop:
Depend more on the event/verb.





Testing Manner/Result tests: Results by diagnostic

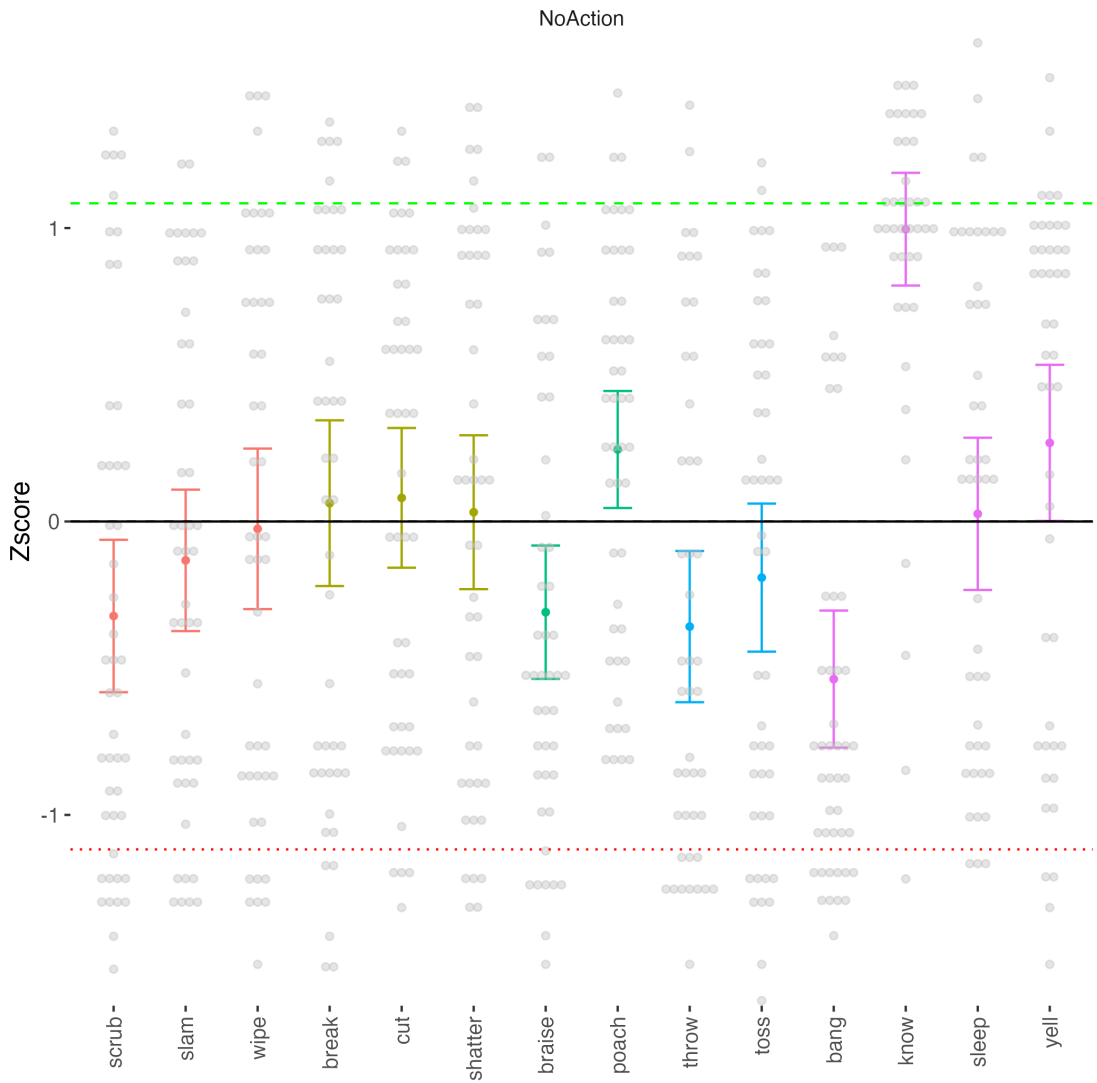
M Chris scrubbed the tub for hours, but didn't move a muscle. Rather, she didn't stop her toddler from strapping scourers to his feet and walking around the empty tub.

R Kim broke my DVD player, but didn't move a muscle. Rather, when I let her borrow it a disc was spinning in it, and she just let it run until the rotor gave out.

C Jessie braised the chard, but didn't move a muscle. Rather, he left the pan on the hob even though the off switch was broken.

T Angus threw Riley the tin, but didn't move a muscle. Rather, he placed in the path of a bouncing basketball, which knocked it forwards.

O Ray banged the drum, but didn't move a muscle. Rather, he let his kids hit it with a stick.



Testing Manner/Result tests: Results by diagnostic

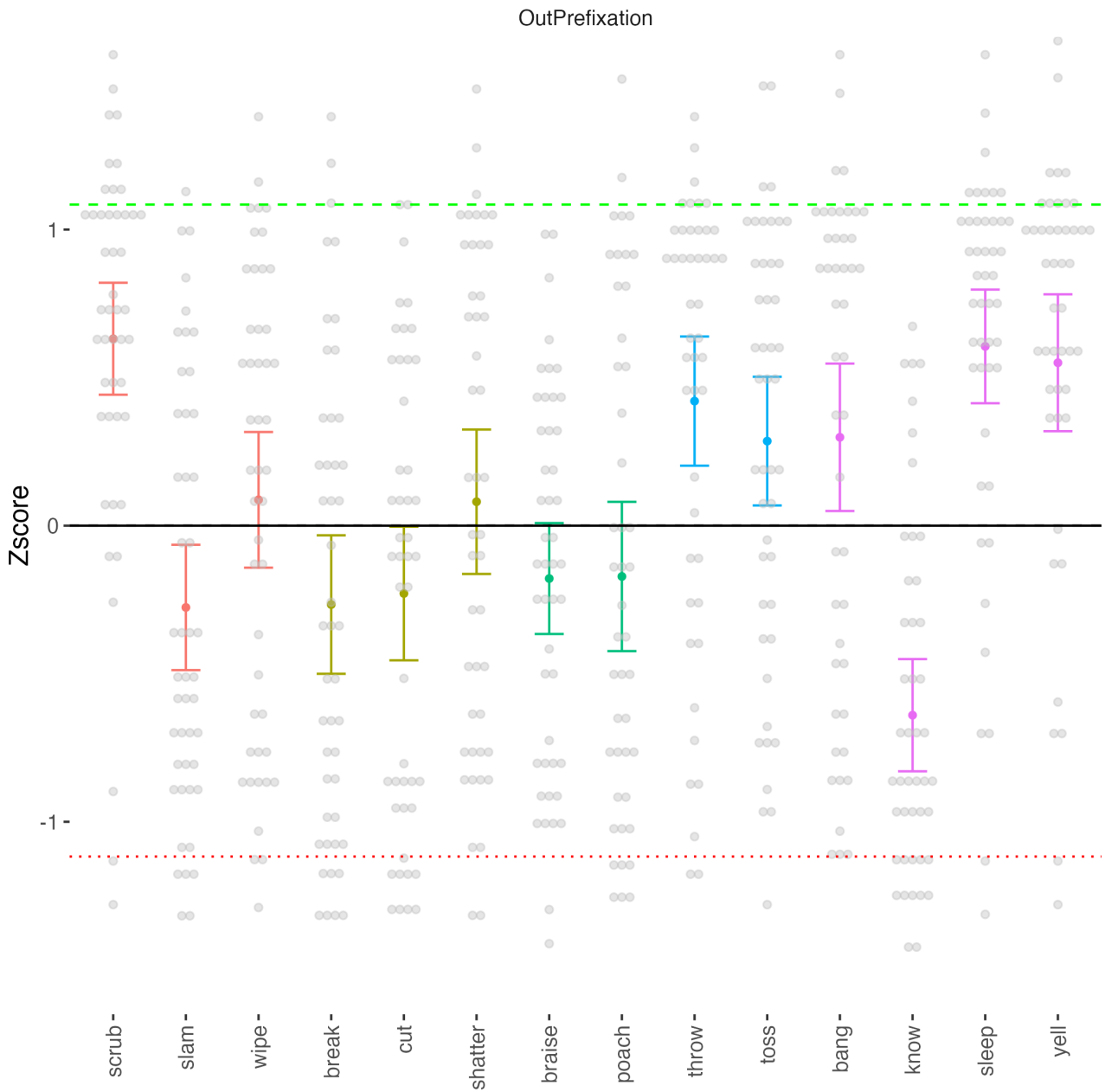
M Chris outscrubbed the other cleaner.

R Kim outbroke the other vase-smasher.

C Jessie outbraised the other chef.

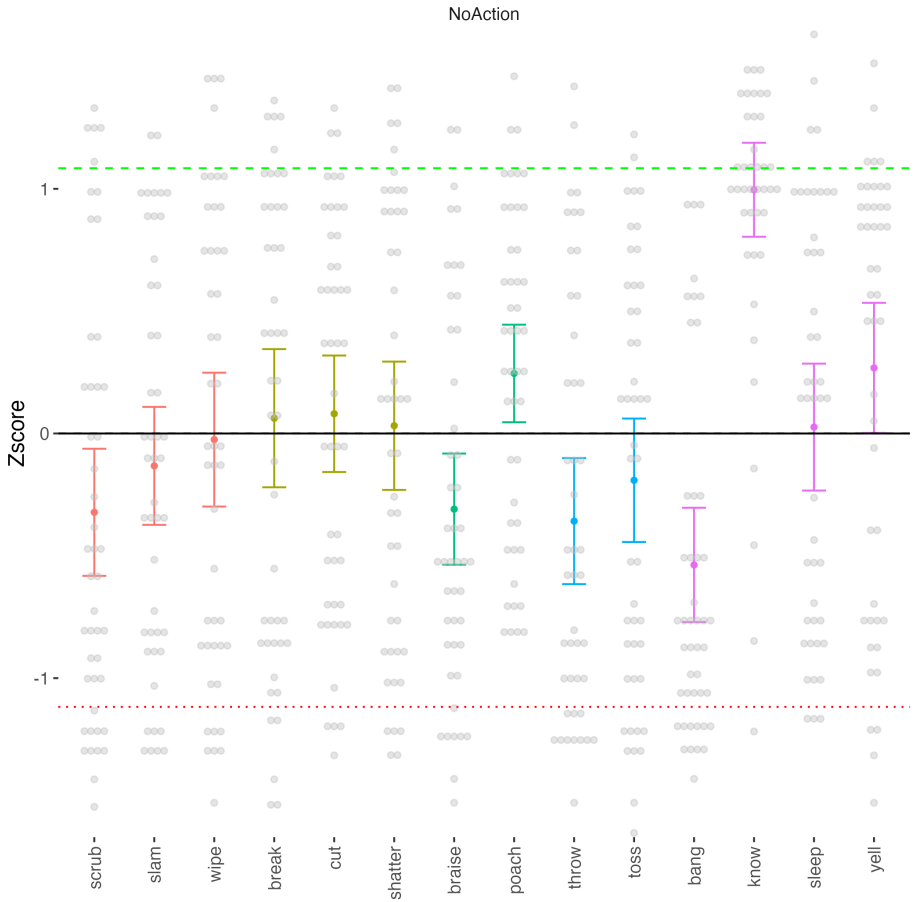
T Angus outthrew the other bowler.

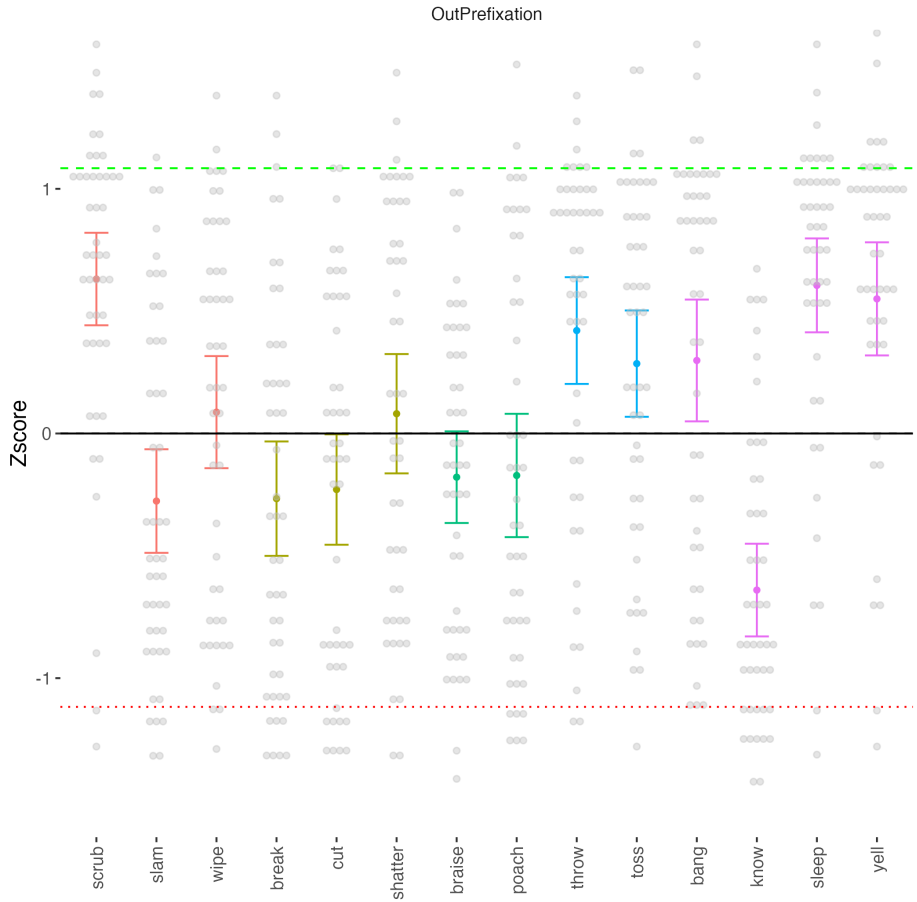
O Ray outbanged the other drummers.



Testing Manner/Result tests: Results by diagnostic

No Action and *Out*-prefixation:
Not particularly good discriminators.
Again pragmatics doing the heavy lifting?





Comparison with the literature

How well do judgments in the literature predict our findings?

$^*, \# \Rightarrow -1$

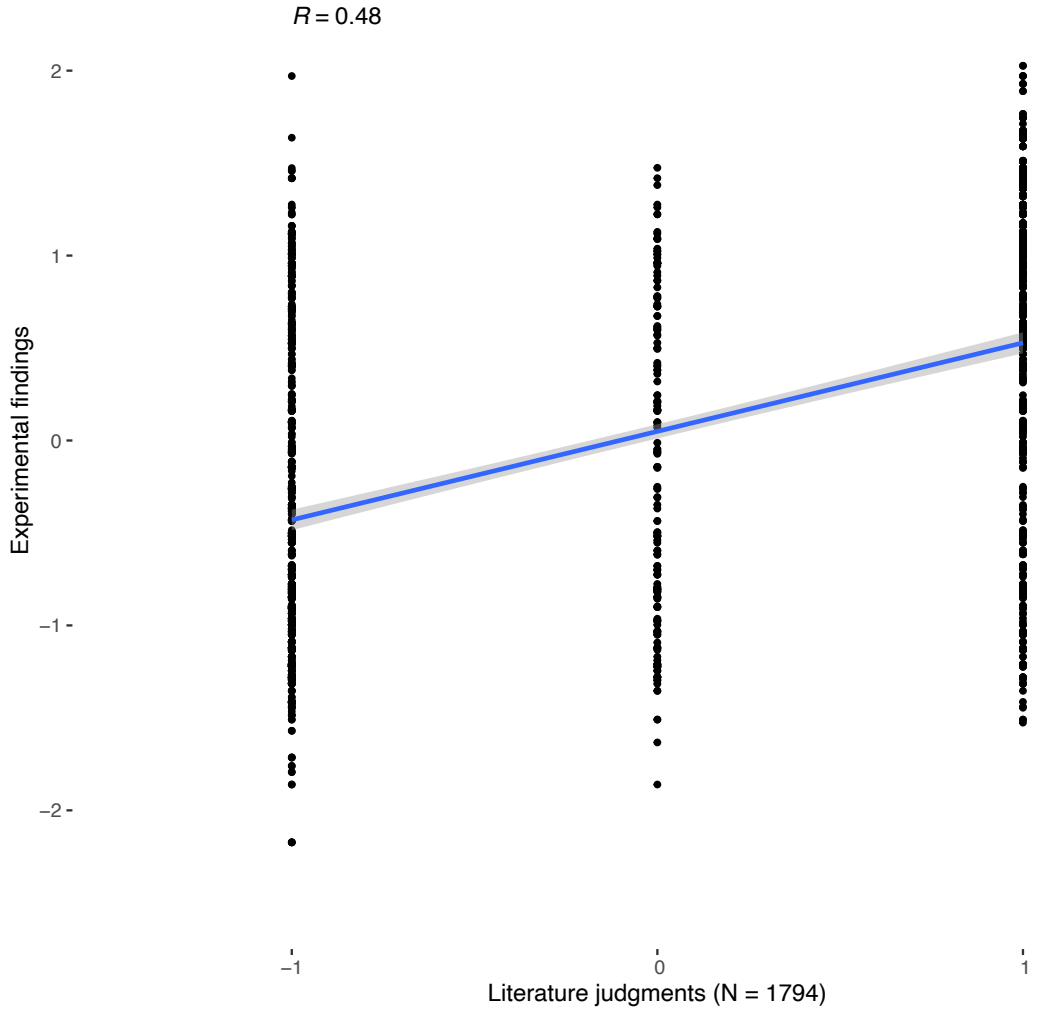
$? \Rightarrow 0$

$\text{ok} \Rightarrow 1$

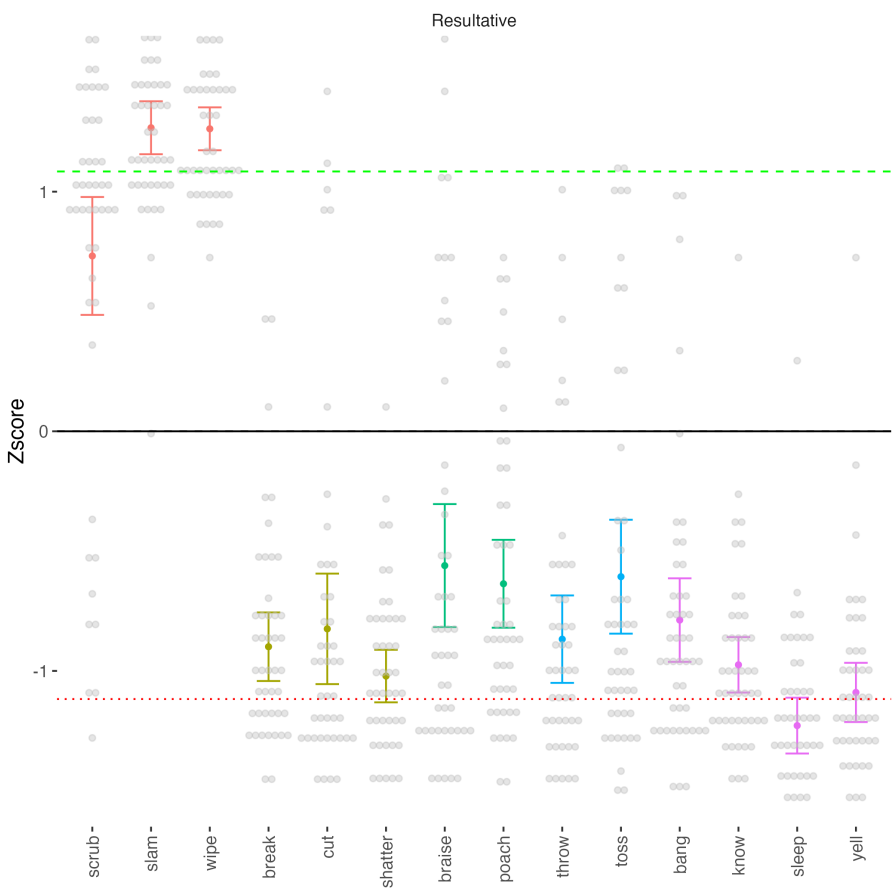
The judgment goes in the same direction (+/−) only in 1171/1794 cases.

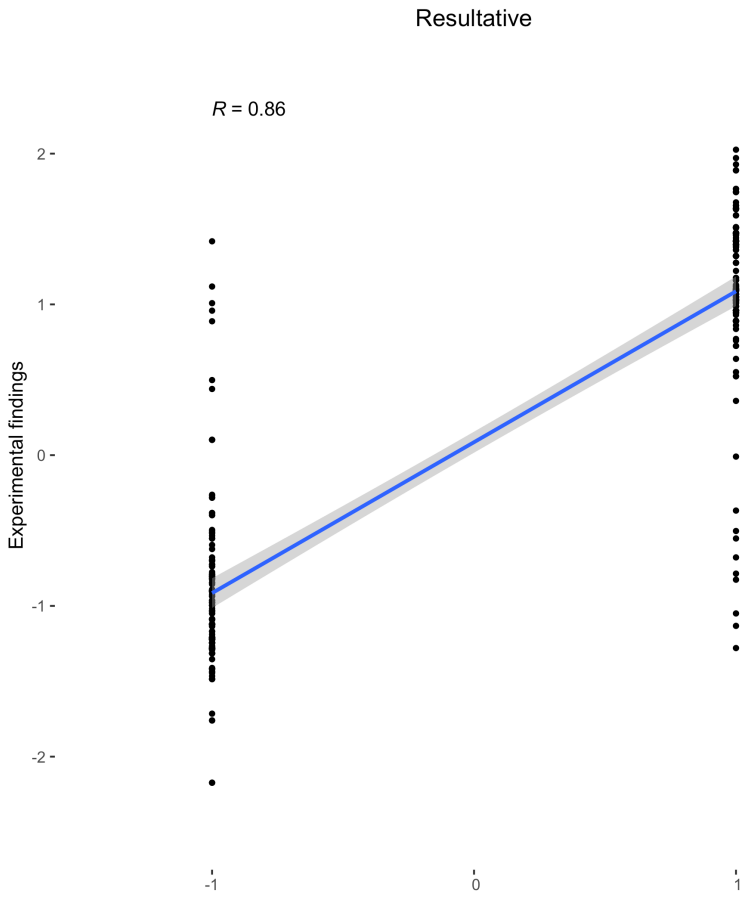
- 65%
- t(3586) = 1.98
- p = 0.048

(cf. Sprouse et al. 2013)



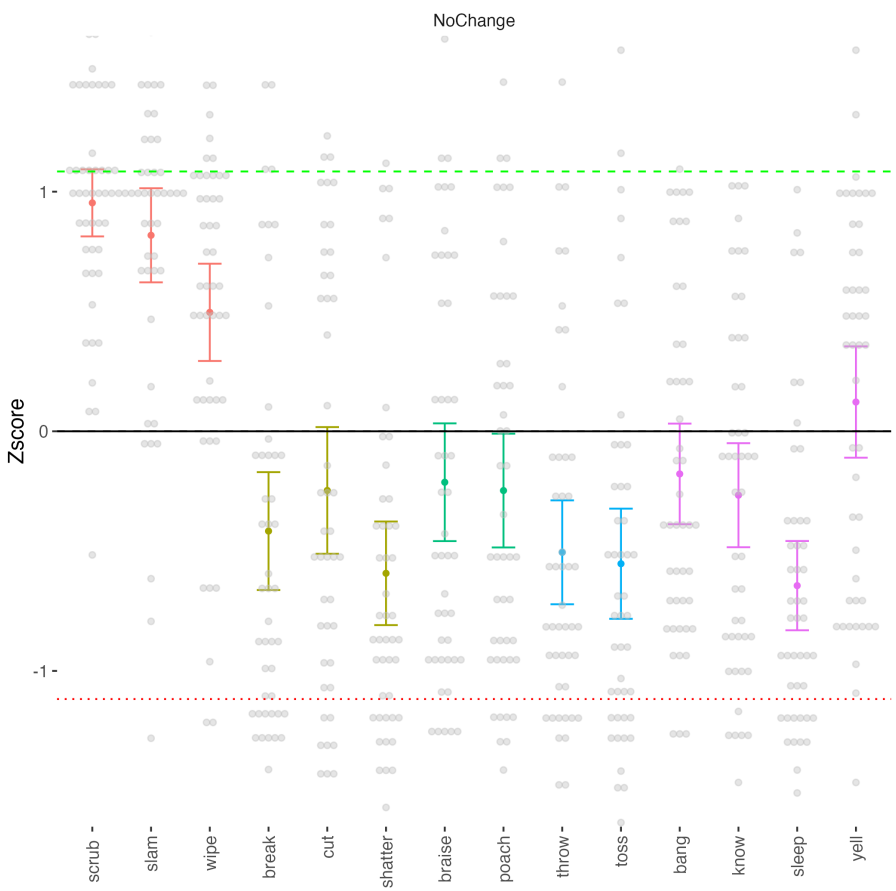
Comparison with the literature

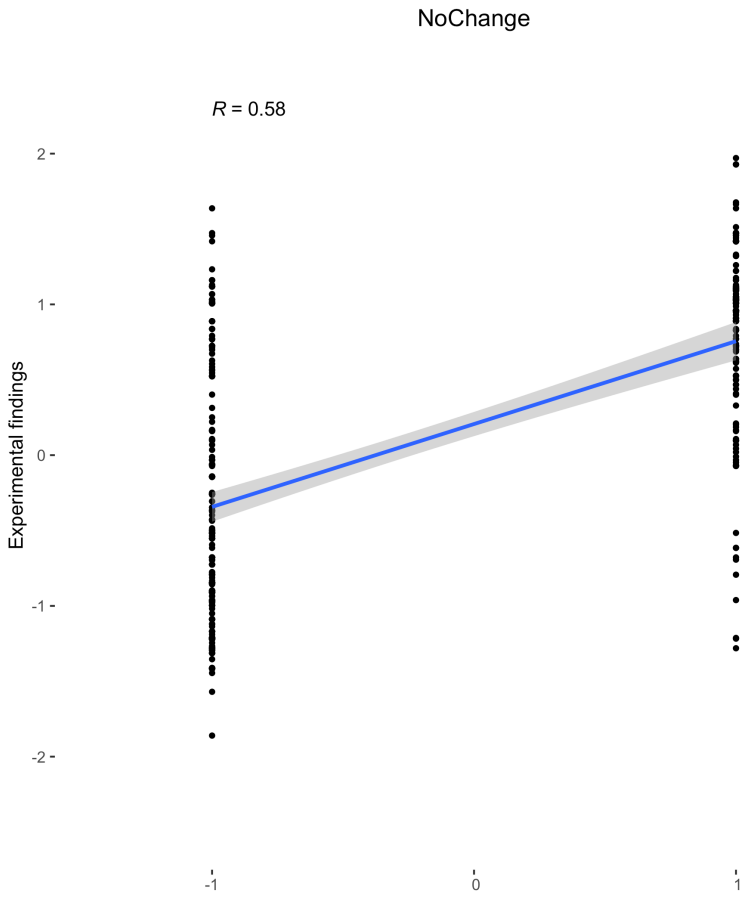




- This looks like a syntactic constraint.
- Impressive that participants got the intended reading!

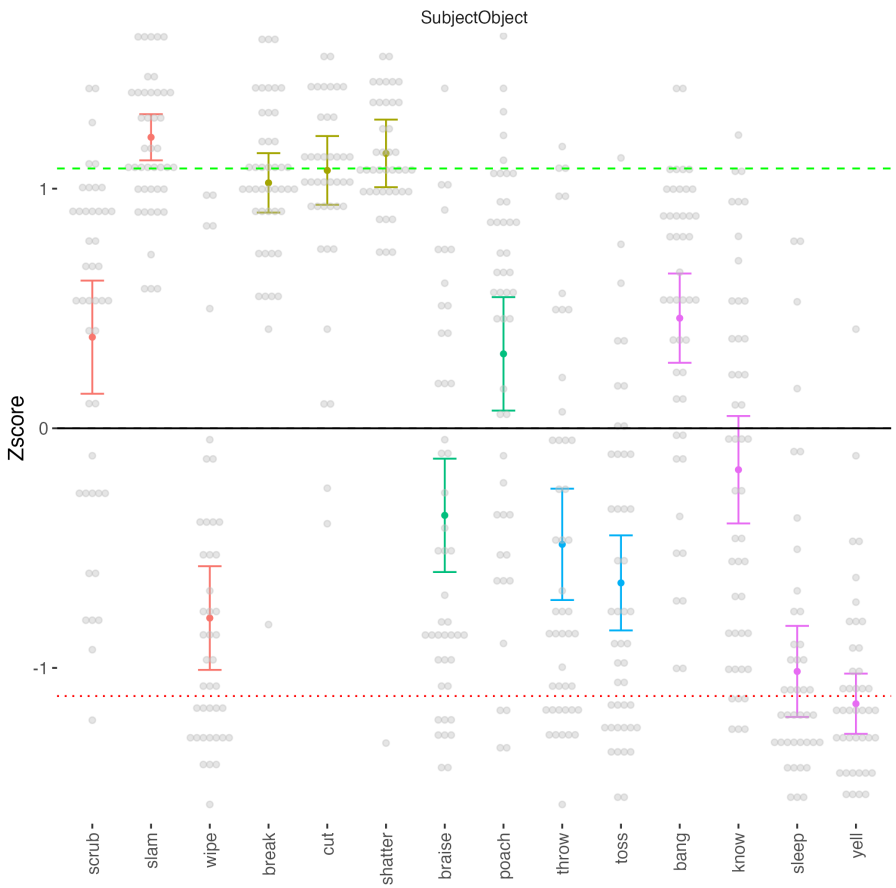
Comparison with the literature

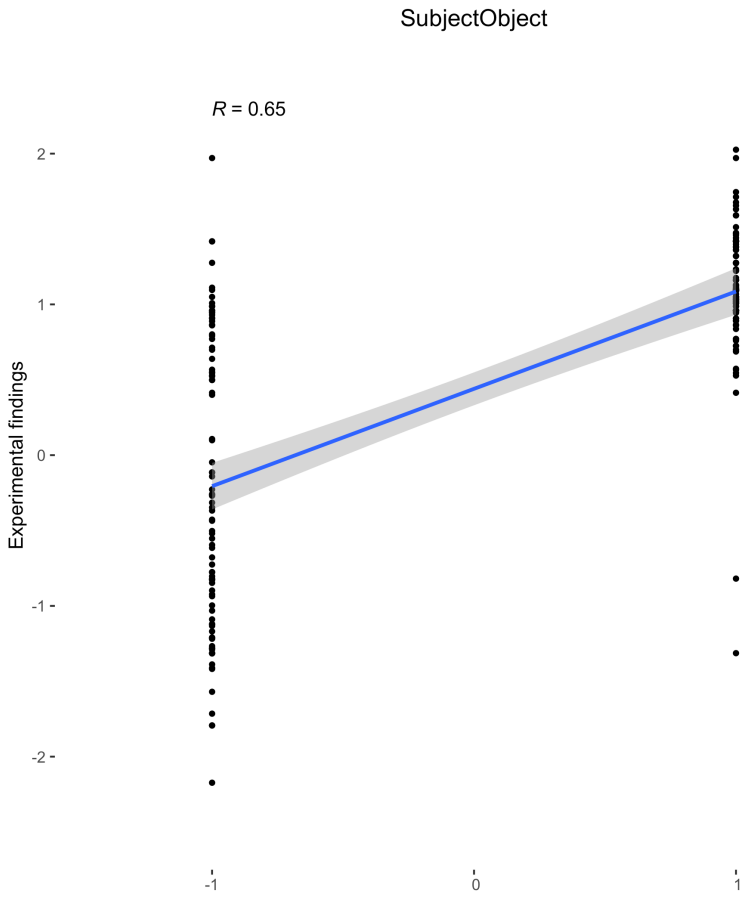




- Directly targets the change pragmatically.
- Fairly robust, but susceptible to context.

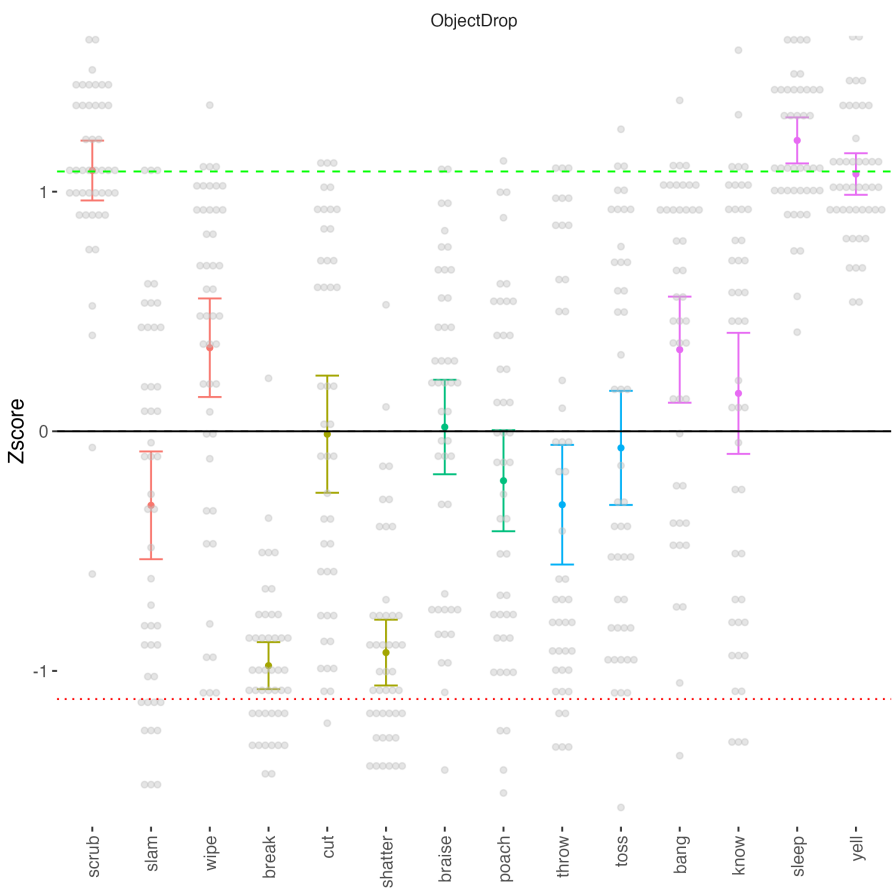
Comparison with the literature

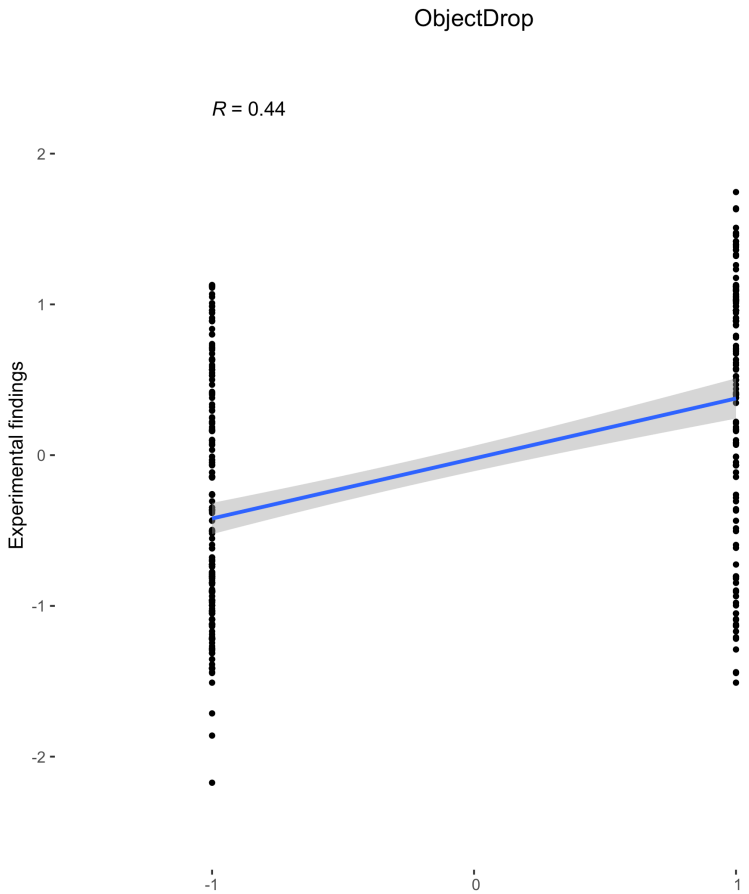




- Depends on the event (predicate).
- Scrubbing odd, slamming fine, breaking fine, yelling terrible.

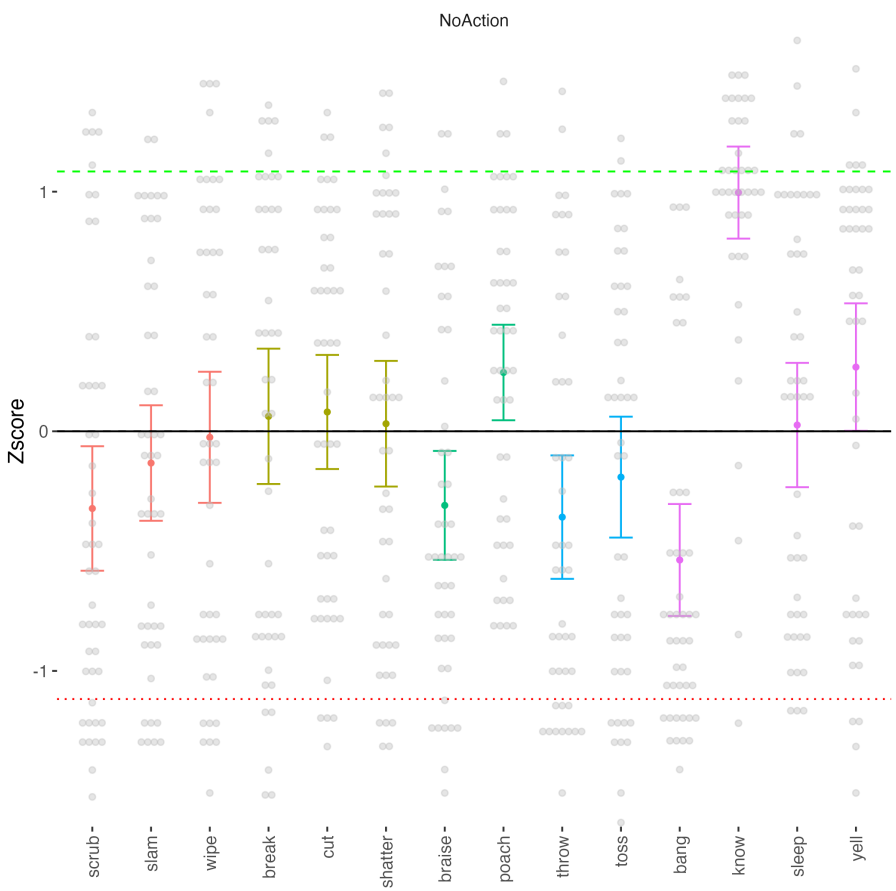
Comparison with the literature

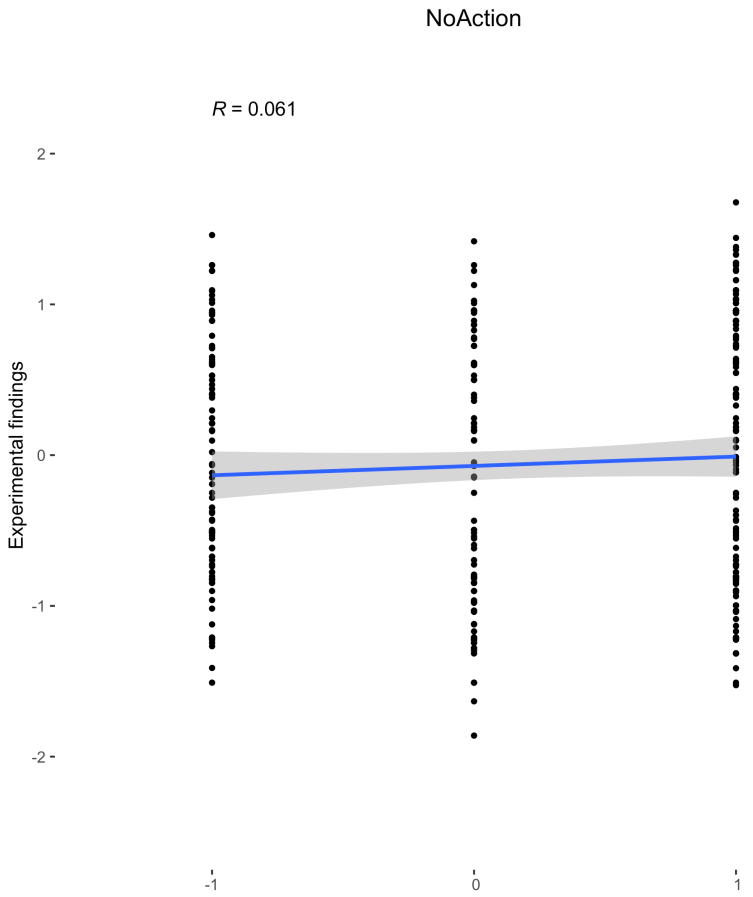




- Depends on the event (predicate) and the prototypical action. (Glass 2021)
- Scrubbing fine, slamming odd, breaking odd, yelling perfect.

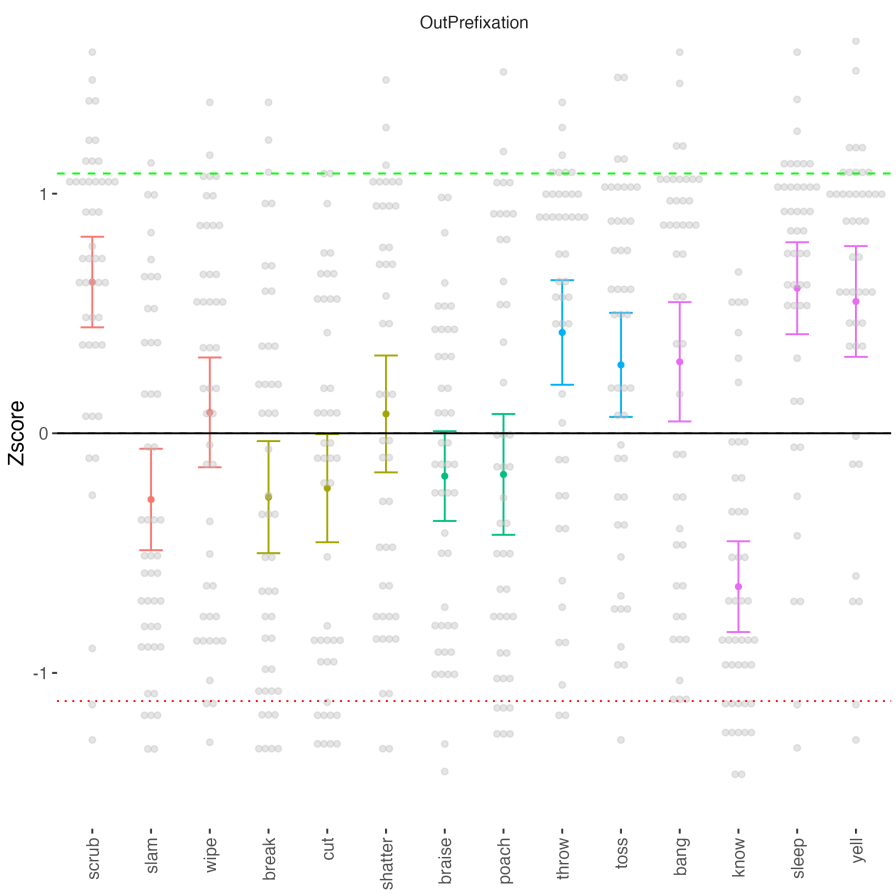
Comparison with the literature

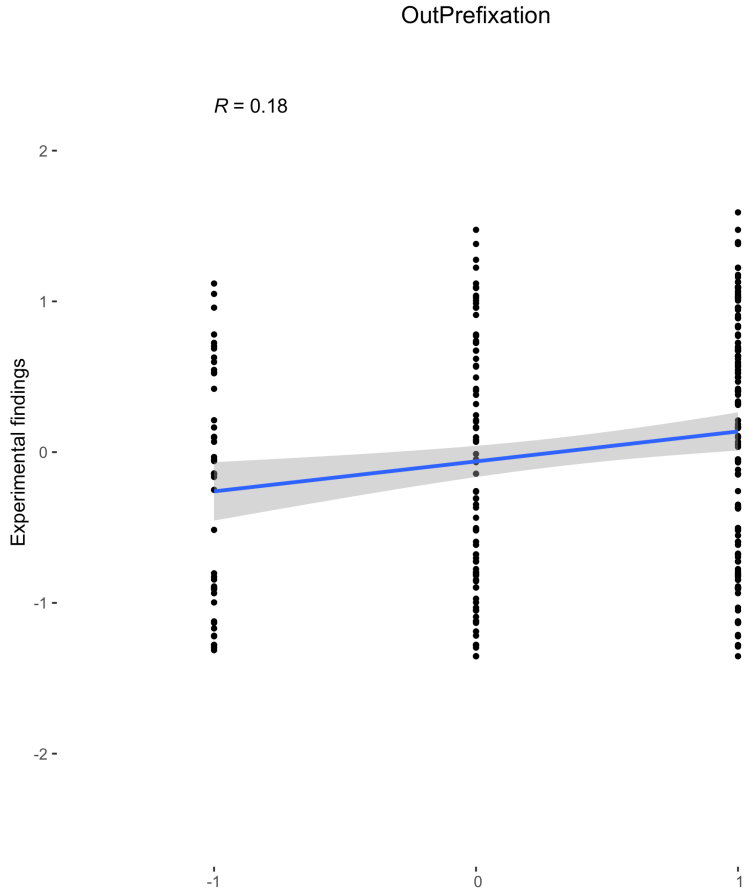




- Lots of pragmatic accommodation.
- States (like knowing) are fine.

Comparison with the literature





- Not reliable in our sample.
- Combination of marked construction and lots of accommodation.

Testing Manner/Result tests: Discussion

Our pilot study

- Tested six Manner/Result diagnostics in an acceptability study.
- Resultatives and denied change are the most robust.
- Perhaps the first study that lets us evaluate syntactic, semantic, pragmatic and lexical aspects of standard diagnostics.

Reminder of the bigger questions:

1. Is Manner/Result the ontology itself? Is the ontology about scales?
2. Are Cooking/Throwing/Killing/Stealing defined at the same level?
3. What about any other verb class?
4. How can we tell what class a given verb is in?

Discussion

1. To our knowledge, the first attempt to control for variation within and across diagnostics.
2. Not all diagnostics in differentiated the two verb classes equally well.
3. What might the diagnostics be probing:
   - Resultatives: syntax.
   - No Action: the ability of pragmatics to make lots of contexts everything sound ok.
   - No Change: pragmatics plus semantics, depending on the verb?
4. Non-target readings in isolation:
   - *John poached all morning long.*
   - *Ray banged the drum torn.*
   - What are we testing when presenting these in isolation (to linguists/participants)?
5. Manner/Result complementarity seems to be less about a grammatical binary and more about different components of meaning (and potentially grammar) that a given context might interact with.
6. Starting point for more tests, more verbs, more verb classes, more languages, ...

Word embeddings: Background

"Word embeddings", "vector space representations":

- Calculate co-occurrence of words and contexts (other words).

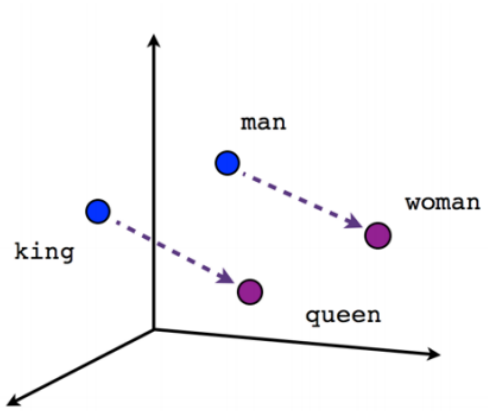
	bite	*buy*	*drive*	*eat*	*get*	*live*	*park*	*ride*	*tell*
bike	0	9	0	0	12	0	8	6	0
car	0	13	8	0	15	0	5	0	0
dog	0	0	0	9	10	7	0	0	1
lion	6	0	0	1	8	3	0	0	0

(Lenci 2018)

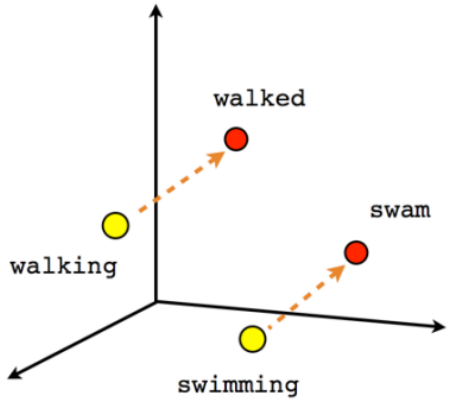
- Create an abstract (vector space) representation of words in a corpus.
- We get an abstract, numerical representation of each word: a vector.
  dog = [2.972568, -0.76399034, 1.3605528, -2.036042, -2.3865438, ...]

Word embeddings: Background

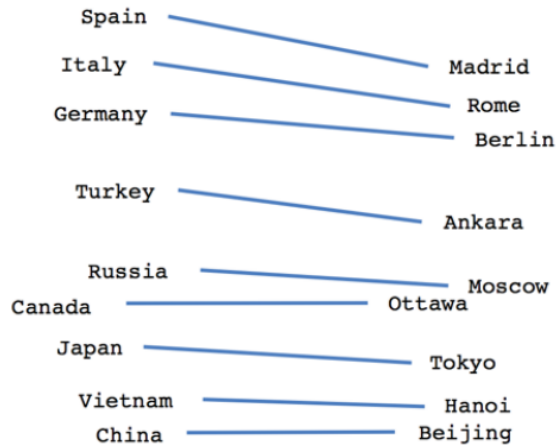
- We get an abstract, numerical representation of each word: a vector.
  dog = [2.972568, -0.76399034, 1.3605528, -2.036042, -2.3865438, …]



Male-Female



Verb tense



Country-Capital

(But cf. Linzen 2016)

- Two common ways of interpreting results:
  - *Reduce the dimensions* from 200 to 2 or 3, and evaluate clusters visually.
  - Calculate quantitative measures.
- As well as downstream tasks (using these embeddings for machine translation, speech recognition, etc).

Word embeddings: Background

Typical questions:

1. What are these models actually learning?
2. How can they be improved (computational advancements)?
3. How can they be improved (linguistic knowledge)?
4. How can they be used for downstream applications?
5. How can be used for theorizing?
6. Do they mirror human performance?
7. Do they mirror acquisition? (Landauer and Dumais 1997)

- Supervised learning: can a simple classifier trained onlabelled data learn to correctly classify verb embeddings as Manner or Results?
- Unsupervised learning: do the embeddings naturally cluster consistently with Manner/Result complementarity?

Word embeddings: Methods

	Manner			**Result**	
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	

Word embeddings: Methods

For the items:

- It can be tricky deciding whether a given verb/root is Manner or Result.
- Used the existing examples in Rappaport Hovav and Levin (2010) and Rappaport Hovav (2017).
- Used only the citation form (past tense singular).
- 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.

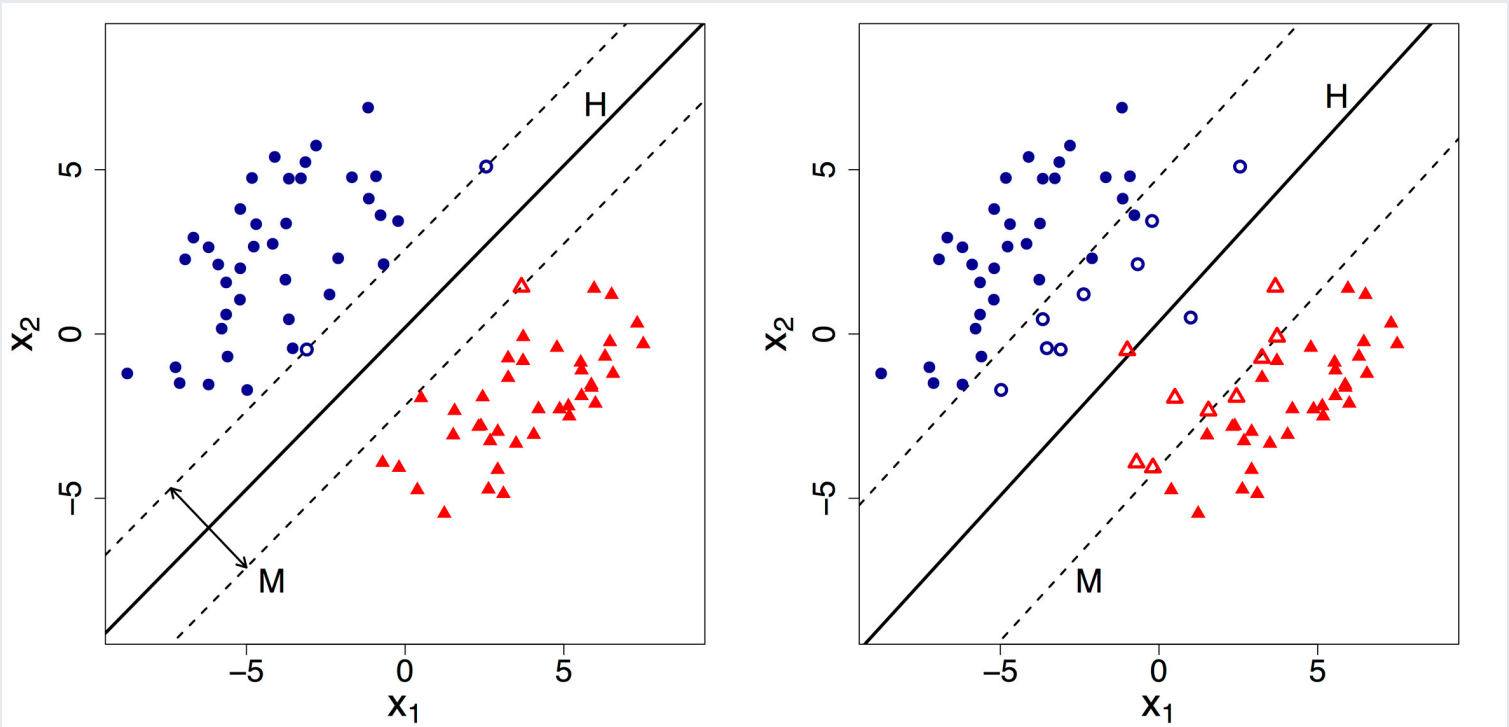
For the model:

- word2vec with 300 dimensions.
- Also a version with syntactic dependency parsing (Levy and Goldberg 2014).

Word embeddings: Methods

Supervised learning: real classification is used to inform learning

Support vector machines as classifiers (schematic figures):



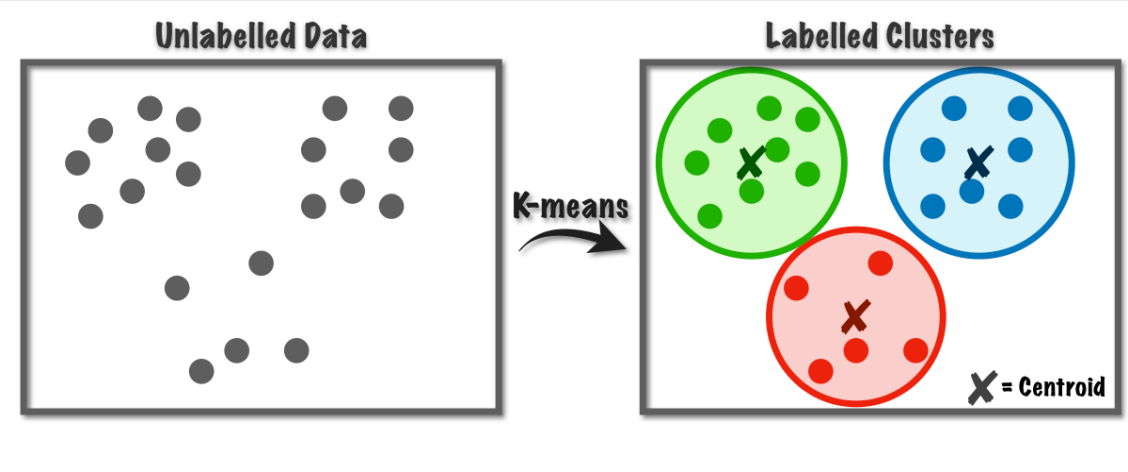
Trained on 75% of the data, tested on the remaining 25%.

➤ 2D visualizations, but multi-dimensional data.

Word embeddings: Methods

Unsupervised learning: not informed by target classification

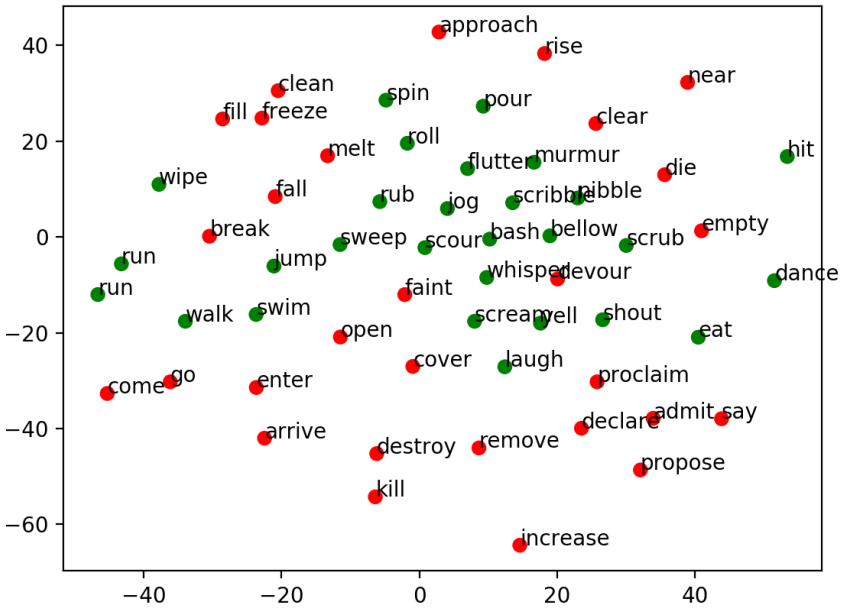
K-means clustering:

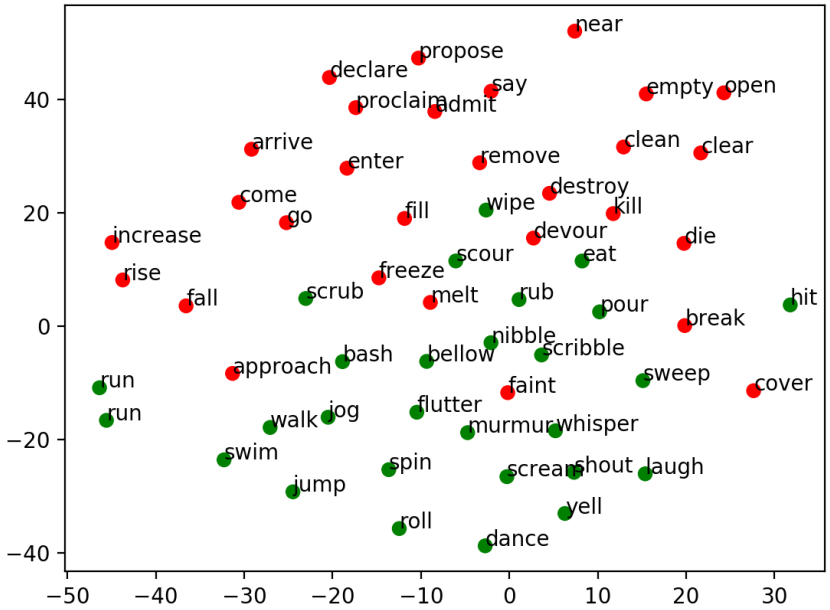


➤ 2D visualizations, but multi-dimensional data.

Word embeddings: Results

2D visualizations (t-SNE dimensionality reduction), for bag-of-words and dependencies:



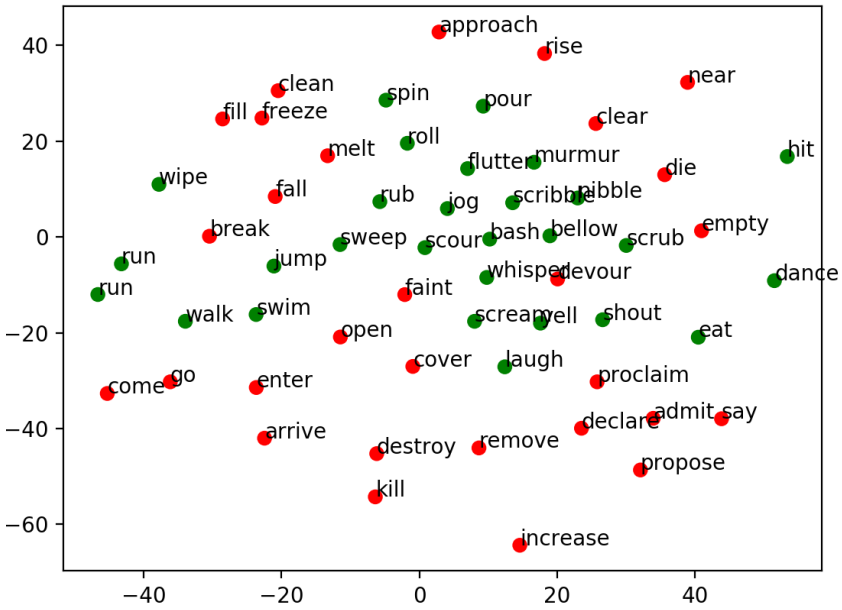


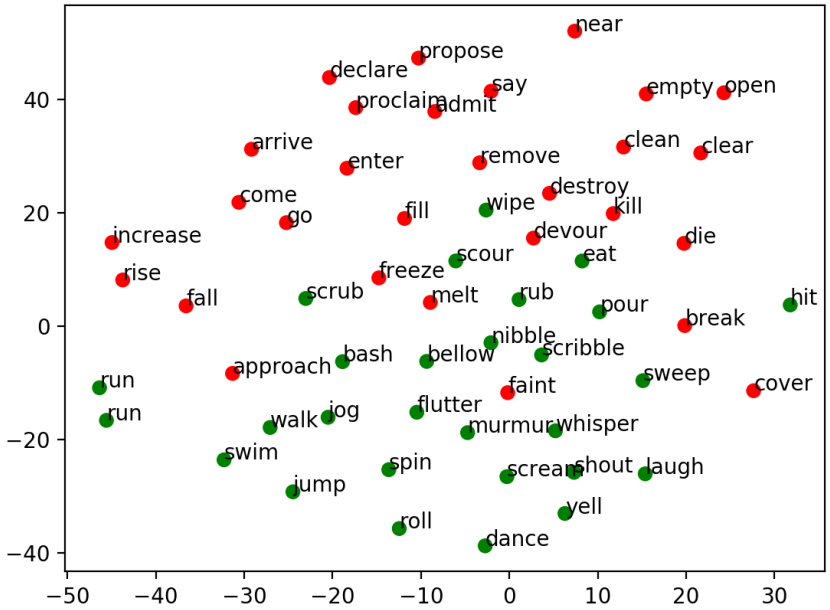
Mean accuracy:

Baselines		SVM		K-means	
Random	Majority	BOW	Dependencies	BOW	Dependencies
57.1%	51.8%	**85.7%**	79.3%	78.1%	77.5%

Word embeddings: Results

2D visualizations (t-SNE dimensionality reduction), for bag-of-words and dependencies:



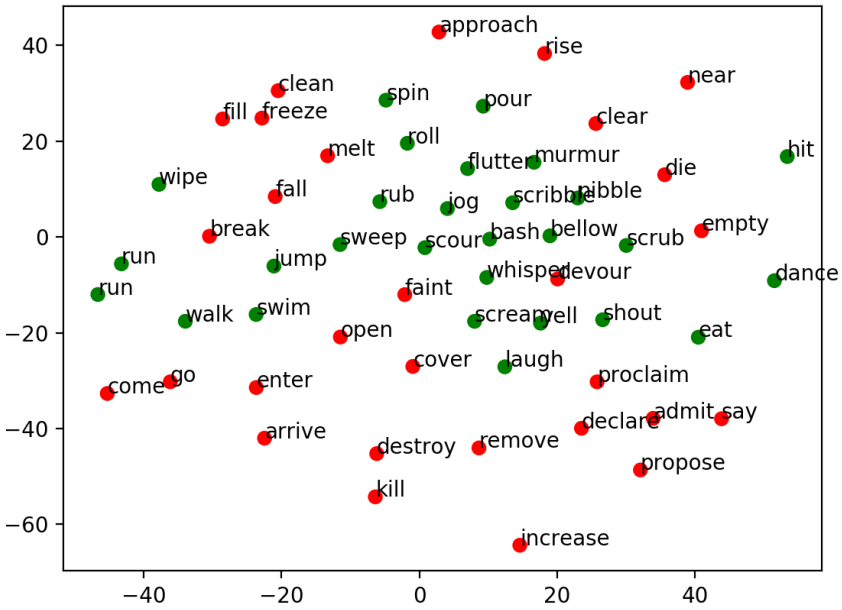


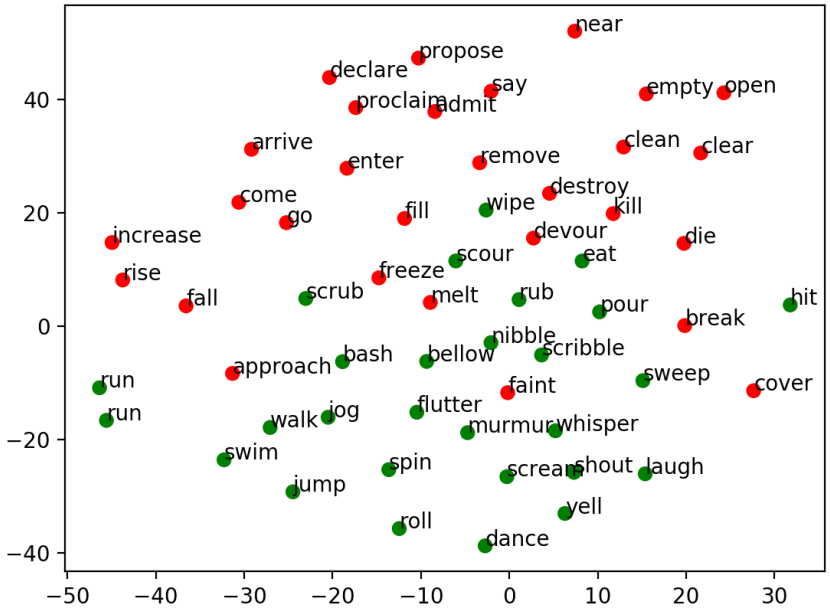
Mean accuracy:

Baselines		SVM		K-means	
Random	Majority	BOW	Dependencies	BOW	Dependencies
57.1%	51.8%	**85.7%**	79.3%	78.1%	77.5%

Word embeddings: Results

2D visualizations (t-SNE dimensionality reduction), for bag-of-words and dependencies:





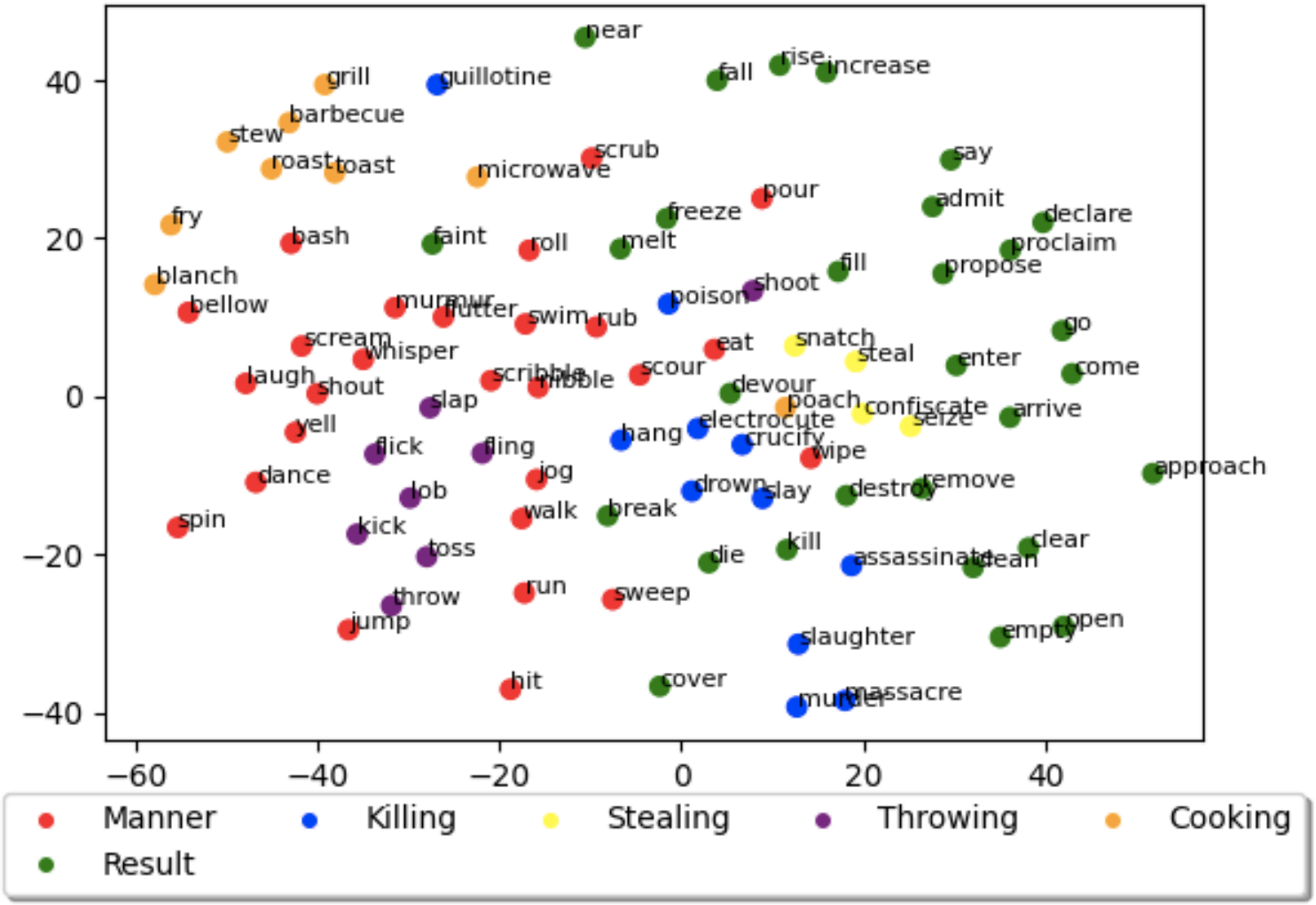
Mean accuracy:

Baselines		SVM		K-means	
Random	Majority	BOW	Dependencies	BOW	Dependencies
57.1%	51.8%	85.7%	79.3%	78.1%	77.5%

Word embeddings: Discussion

- A simple word embedding model captures Manner/Result Complementarity.
- But is it good for the linguists?
- Recall again that some roots/verbs have been argued to encode both Manner and Result.
  - Verbs of stealing: *steal, rob, snatch, seize, confiscate, ...*
  - Verbs of cooking: *poach, roast, sautee, braise, ...*
  - Verbs of directed throwing: *throw, kick, toss, flip, fling, ...*
  - Verbs of killing: *massacre, slay, crucify, drown, hang, ...*

Word embeddings: Discussion



- Progress towards quantitative evaluation of empirical claims.
- Developing quantitative measures of closeness.

1. Introduction
2. Testing Manner/Result tests
   - Background
   - Methods
   - Results by diagnostic
   - Comparison with the literature
   - Discussion
3. Word embeddings
   - Background
   - Methods
   - Results
   - Discussion
4. LLMs
   - Surprisal
   - Probing
   - Summary
5. Conclusion

LLMs

Some limitations of word embeddings

- Static word embeddings aggregate word information across all contexts.
- Lexical semantic properties might get diluted.
- Possible solution #1: sense-ful embeddings (Eyal et al. 2022).
  - ⇒ Back-up slides. Ultimately not that much better.
- Possible solution #2: transformers (Large Language Models, LLMs).

- *Surprisal*: how "surprised" a Large Language Model is when it encounters an unexpected token.

(Hale 2006; Linzen and Jaeger 2015; Wilcox et al. 2024)

- *Probing*: figuring out what happens inside the different layers of an LLM.

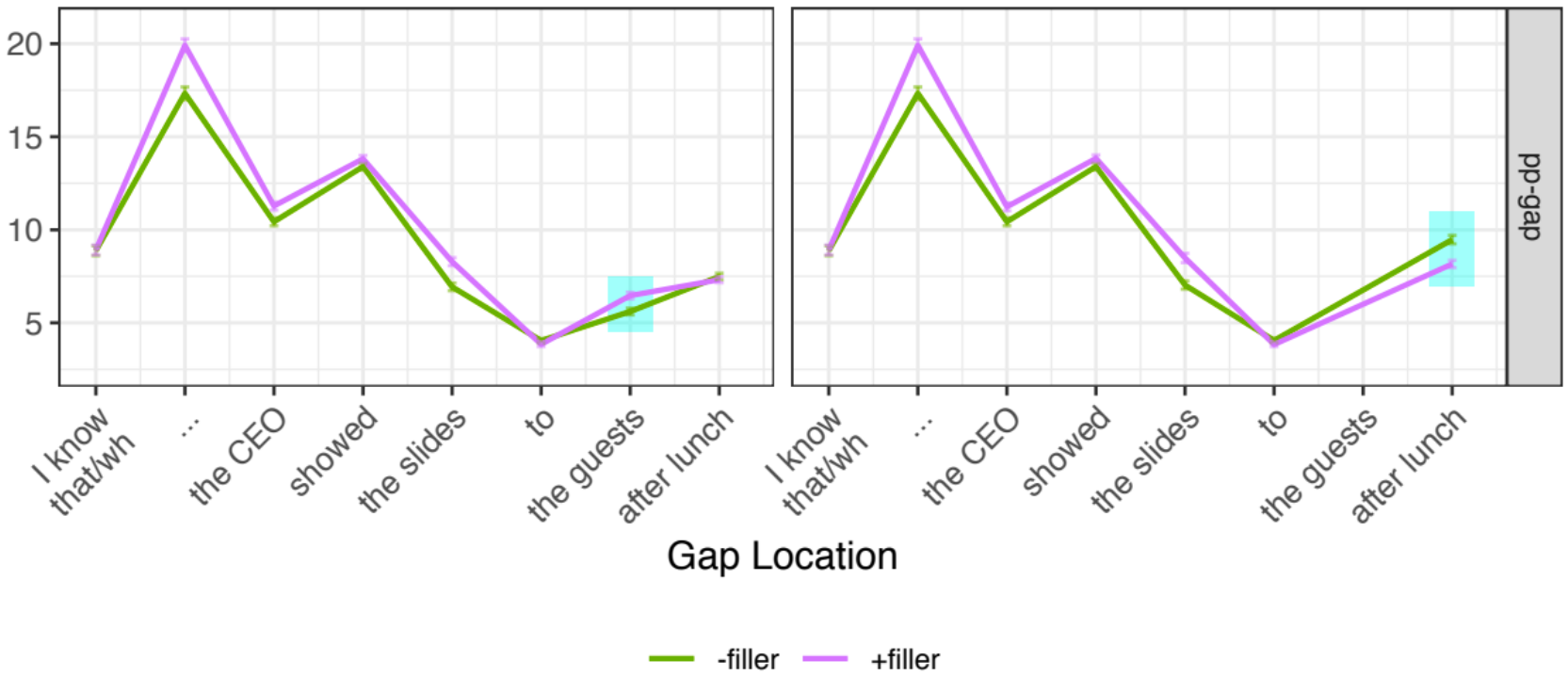
(Clark et al. 2019; Ethayarajh 2019; Reif et al. 2019; Jin et al. 2025)

LLMs: Surprisal

Wilcox et al. (2024): various models have higher surprisal when encountering a syntactic mistake.

(10) I know that/*who the CEO showed the slides to the guests after lunch. [left]

(11) I know *that/who the CEO showed the slides to ___ after lunch. [right]

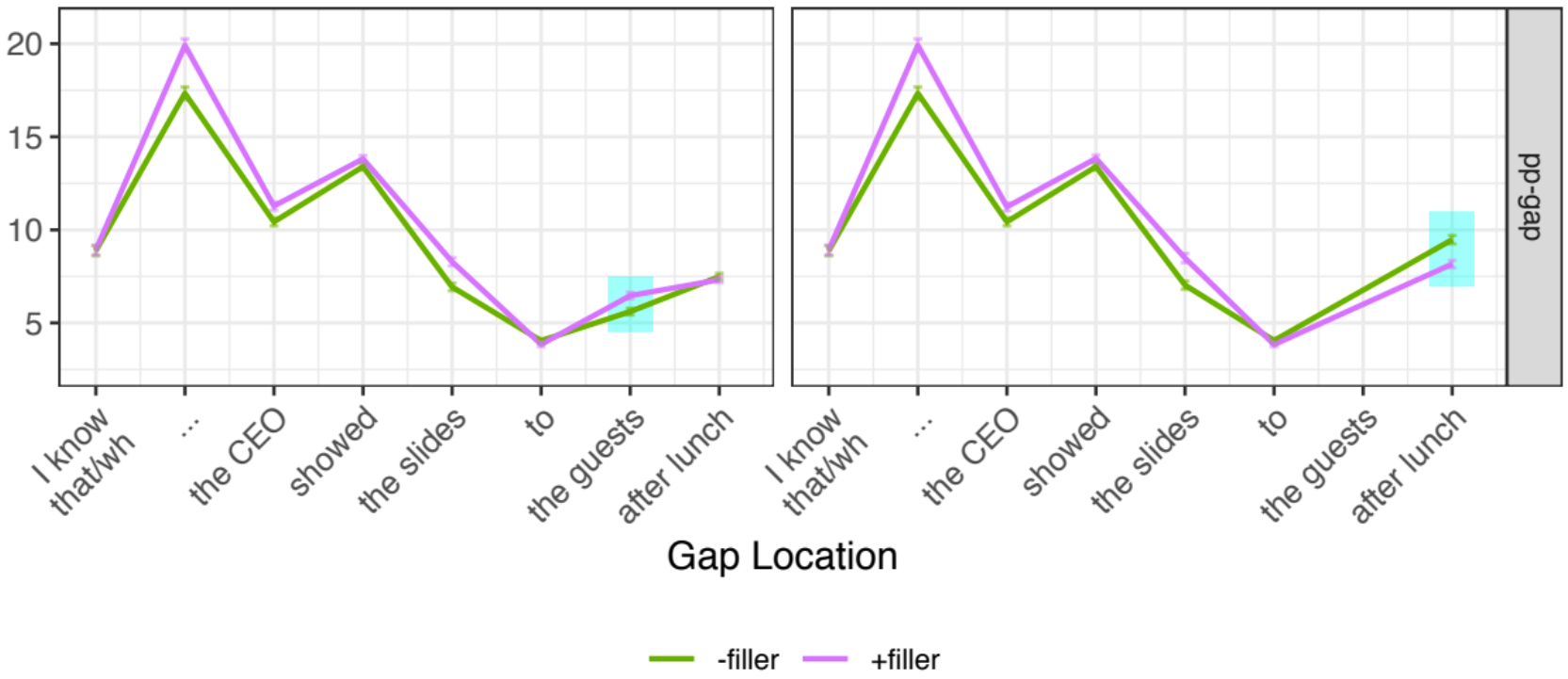


LLMs: Surprisal

Wilcox et al. (2024): various models have higher surprisal when encountering a syntactic mistake.

(10) I know that/*who the CEO showed the slides to the guests after lunch. [left]

(11) I know *that/who the CEO showed the slides to ___ after lunch. [right]



LLMs: Surprisal

Surprisal

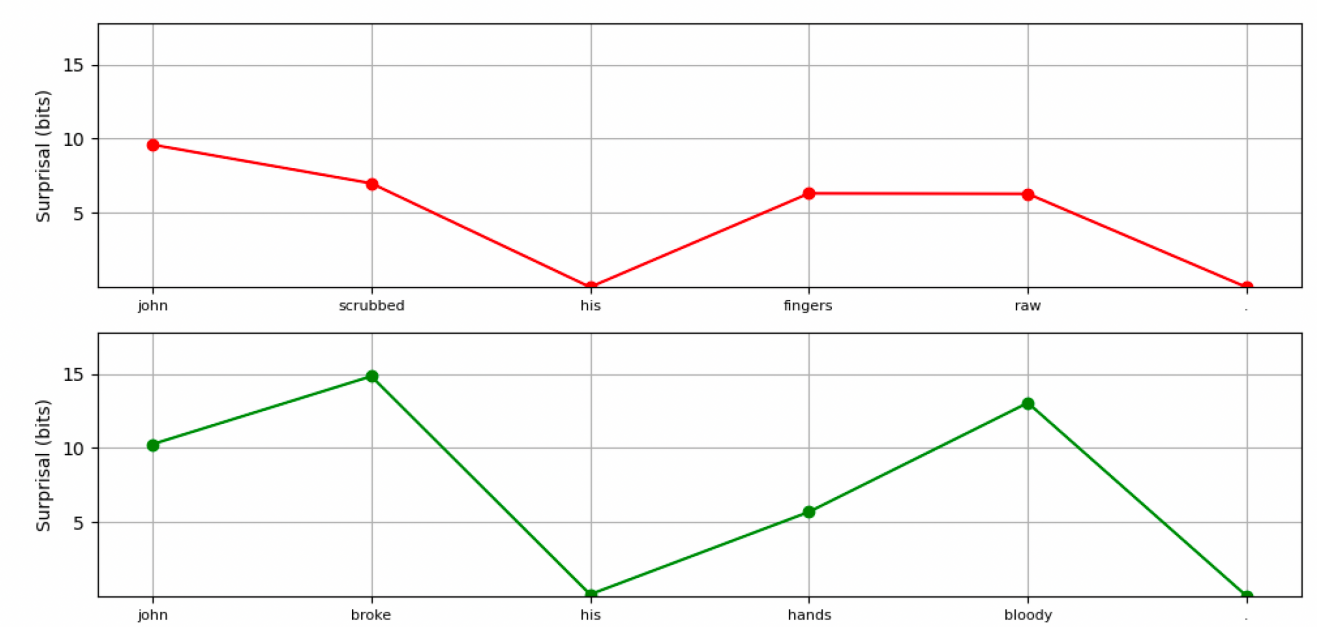
1. Construct sentences following the diagnostics from the literature.
2. At each step, "mask" one item in the sentence.
   - [MASK] *last night, John broke.*
   - *All last* [MASK], *John broke.*
   - ...
3. BERT outputs a probability distribution over tokens at the position of the masked item.
4. Compute surprisal for the word in the original sentence (*night*).
5. More unexpected words get higher surprisal values.

LLMs: Surprisal

- Two Manner verbs (*scrub, walk*), two Result verbs (*break, arrive*), one control (*think*).
- Five diagnostics: **Resultative**, **Denial of Action**, Object Drop, Out-prefixation, Denial of Result.

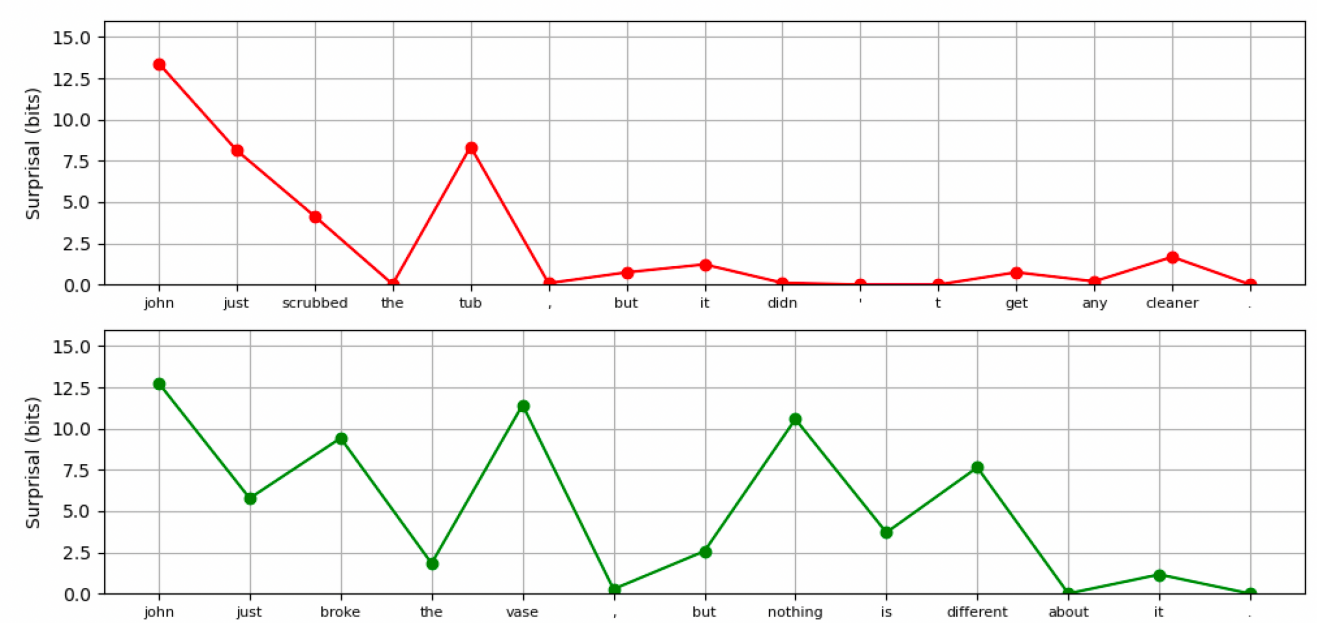
LLMs: Surprisal

Resultatives: as predicted, higher surprisal for the Result verb *broke*.



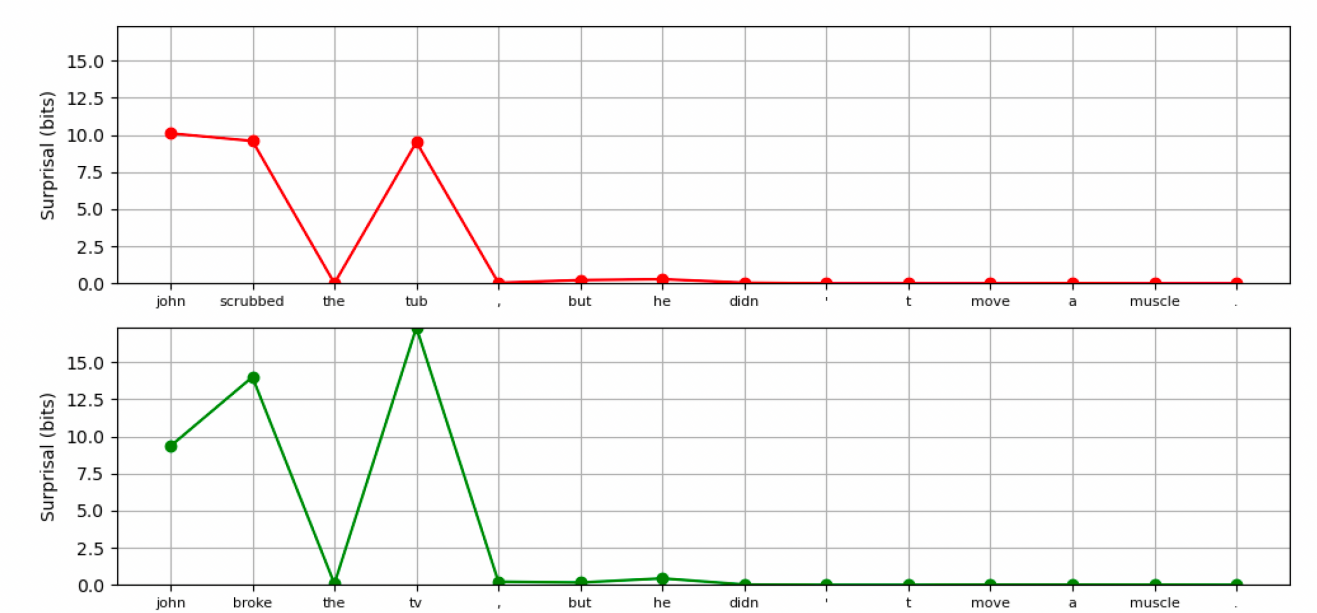
LLMs: Surprisal

No change: as predicted, higher surprisal for the Result verb *broke*.



LLMs: Surprisal

But No Action, shows the opposite of prediction: once again higher surprisal for *break*.



➤ *Break* is also just more frequent than *scrub*.

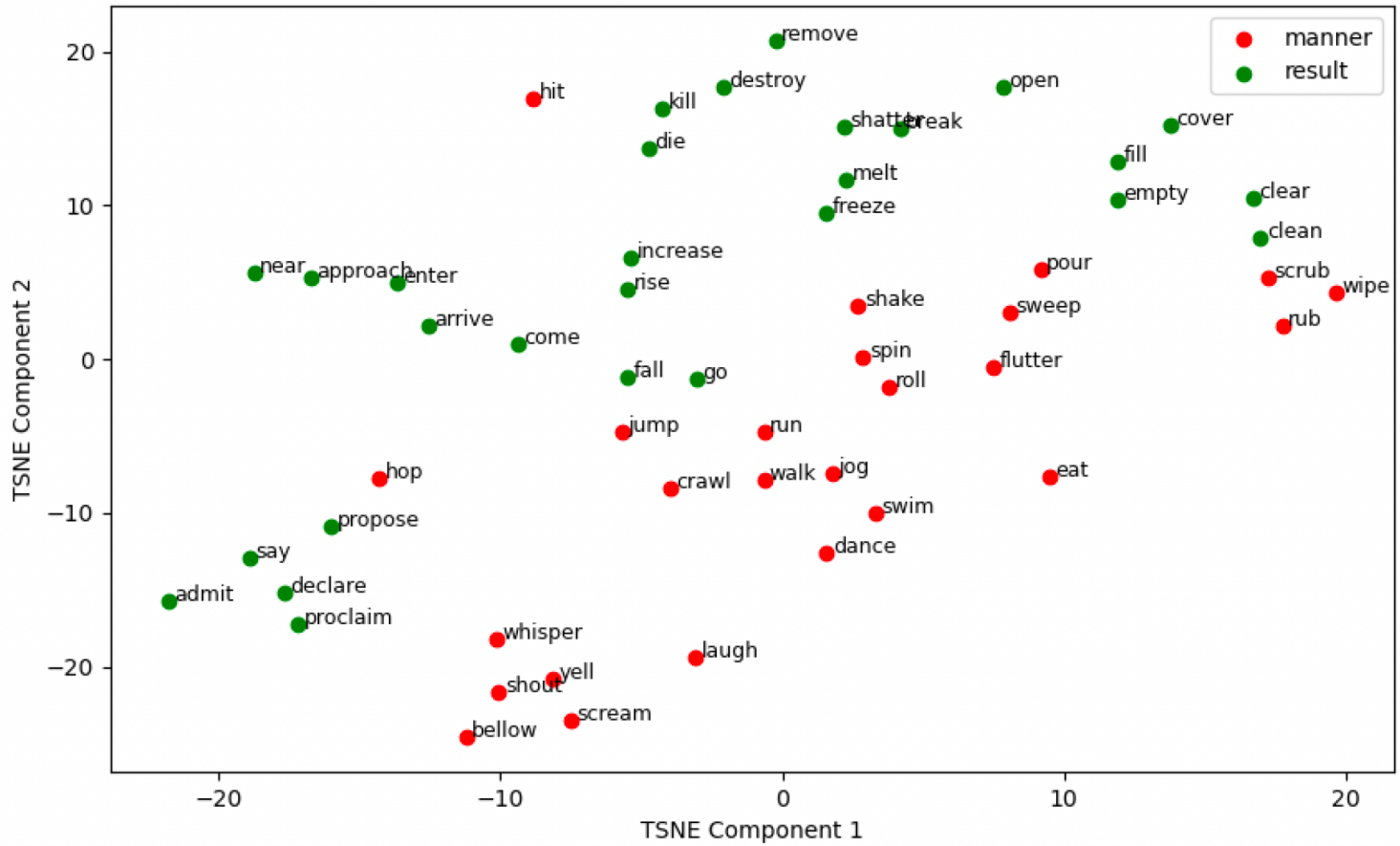
LLMs: Probing

Probing

1. For each verb, select 20 example sentences, controlling for polysemy.
2. Extract verb embeddings across examples and average them to obtain a mean word embedding for each verb.
3. Visualize the embeddings in two dimensions.
4. Perform logistic regression to see if a binary classification can be learned from the embeddings.
5. Repeat for multiple BERT layers.

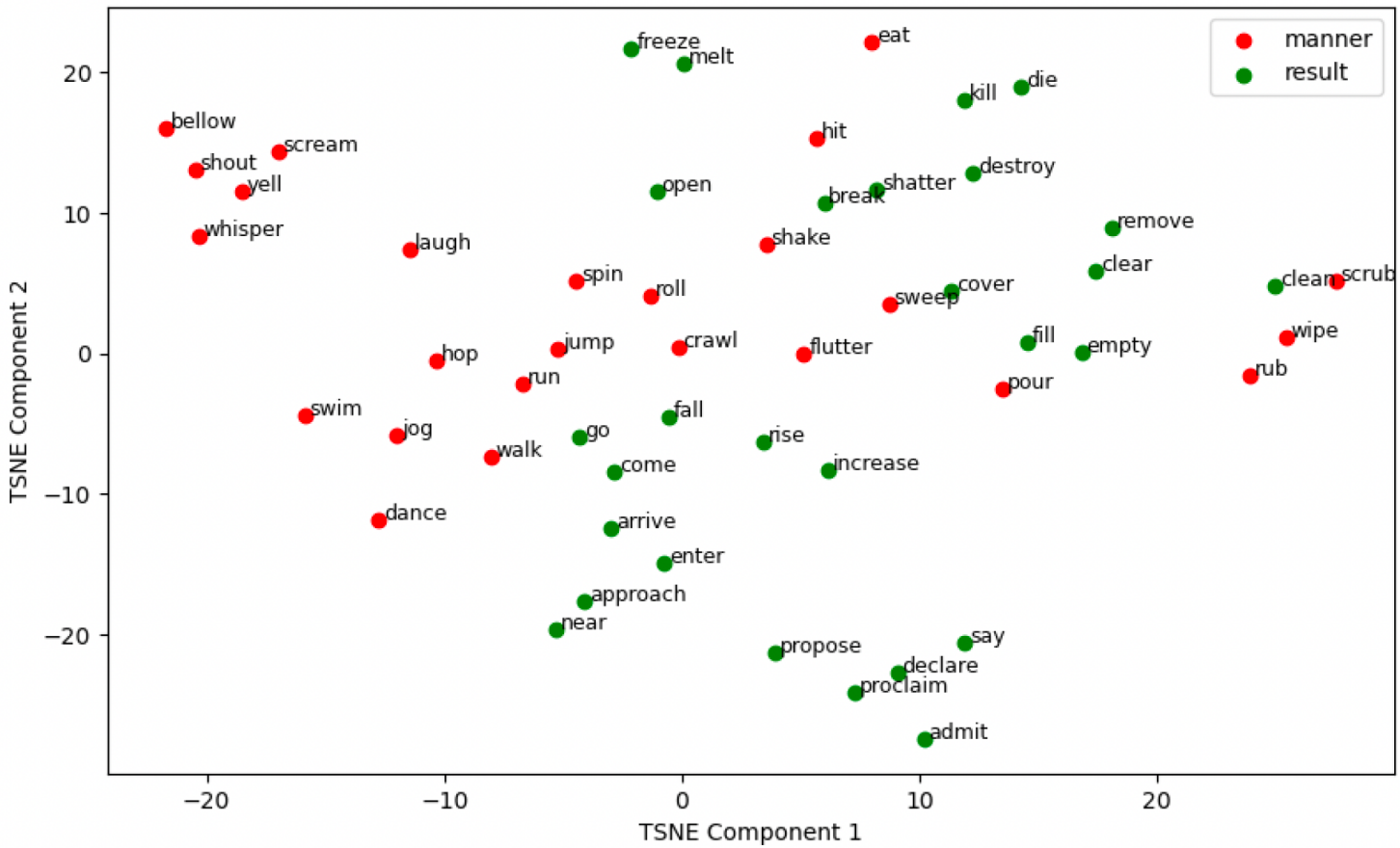
LLMs: Probing

BERT layer 6:



LLMs: Probing

BERT layer 12:



LLMs: Probing

Confirmed in the regression model (5-fold cross validation):

Model (layer)	F-score	AUC-ROC
BERT-large (24)	0.91	0.97
BERT-base (12)	0.88	0.96
BERT-base (6)	**0.95**	**0.99**
BERT-base (2)	0.90	0.97

LLMs: Summary

- Surprisal:
  1. Surprisal seems to conflate grammaticality/acceptability with other factors, e.g. token frequency and conditional probability given context words.
  2. What difference in surprisal is significant?
  3. How should we map surprisal to a measure of acceptability/grammaticality?
- Probing:
  1. BERT layer 6 seems to encode Manner/Result.
  2. BERTology: What else does this layer do? What about other models? What about other verb classes?
- LLMs are sensitive to context in examples – there are lots of possible confounds.
- An interesting question is how this is different from the syntactic/semantic/pragmatic effects on human judgements.

1. Introduction
2. Testing Manner/Result tests
   - Background
   - Methods
   - Results by diagnostic
   - Comparison with the literature
   - Discussion
3. Word embeddings
   - Background
   - Methods
   - Results
   - Discussion
4. LLMs
   - Surprisal
   - Probing
   - Summary
5. Conclusion

Conclusion

What we started with:

What's a verb class?

- At what level is it specified?
- What are the relevant properties?
  - Syntactic
  - Semantic
  - World knowledge
  - Frequency
  - ...
- What's consistent crosslinguistically?
- What are the structural primitives (morphemes/features/functions/operators)?

Conclusion

The literature is ultimately correlational. What about causation?

(12) Expose participants to…
  a. Chris wugged
  b. Chris wugged all day yesterday
  c. Chris wugged the niz halfway
  d. Chris wugged the niz, but nothing is different about it

(13) And then see their judgements on…
  a. The wind blixed the niz
  b. *Kim blixed the niz clean
  c. *Kim blixed all day yesterday

Your thoughts?

Thank you!

- Dan Lassiter and Rob Truswell.
- Paolo Cassina.

Research assistants Leon Cosgrove, Violette Daures, Connor Mathews-Sweetman and Andrew Nixon.

References I

Acedo-Matellán, Víctor, and Jaume Mateu. 2014. From syntax to roots: A syntactic approach to root interpretation. In *The syntax of roots and the roots of syntax*, ed. Artemis Alexiadou, Hagit Borer, and Florian Schäfer, 14–32. Oxford: Oxford University Press.

Alexiadou, Artemis, Fabienne Martin, and Florian 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*.

Ausensi, Josep. 2023. *The division of labor between grammar and the lexicon: An exploration of the syntax and semantics of verbal roots*. Studies in Generative Grammar. Berlin: De Gruyter Mouton.

Beavers, John, and Andrew Koontz-Garboden. 2012. Manner and result in the roots of verbal meaning. *Linguistic Inquiry* 43:331–369.

Beavers, John, and Andrew Koontz-Garboden. 2017. Result verbs, scalar change, and the typology of motion verbs. *Language* 93:842–876.

Beavers, John, and Andrew Koontz-Garboden. 2020. *The roots of verbal meaning*. Oxford Studies in Theoretical Linguistics. Oxford: Oxford University Press.

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

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, ed. Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes, 276–286. Florence, Italy: Association for Computational Linguistics. URL `https://aclanthology.org/W19-4828/`.

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

Ethayarajh, Kawin. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, ed. Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, 55–65. Hong Kong, China: Association for Computational Linguistics. URL `https://aclanthology.org/D19-1006/`.

Eyal, Matan, Shoval Sadde, Hillel Taub-Tabib, and Yoav Goldberg. 2022. Large scale substitution-based word sense induction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4738–4752. Dublin, Ireland: Association for Computational Linguistics. URL `https://aclanthology.org/2022.acl-long.325`.

Glass, Lelia. 2021. English verbs can omit their objects when they describe routines. *English Language and Linguistics* 26:49–73. URL `http://dx.doi.org/10.1017/S1360674321000022`.

Hale, John. 2006. Uncertainty about the rest of the sentence. *Cognitive Science* 30:643–672.

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

References II

Jin, Mingyu, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. 2025. Exploring concept depth: How large language models acquire knowledge and concept at different layers? In *Proceedings of the 31st International Conference on Computational Linguistics*, ed. Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, 558–573. Abu Dhabi, UAE: Association for Computational Linguistics. URL `https://aclanthology.org/2025.coling-main.37/`.

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.

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. 1991. Wiping the slate clean: A lexical semantic exploration. *Cognition* 41:123–151.

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

Levin, Beth, and Malka Rappaport Hovav. 2013. Lexicalized meaning and manner/result complementarity. In *Subatomic semantics of event predicates*, ed. Boban Arsenijević, Berit Gehrke, and Rafel Marín, 49–70. Dordrecht: Springer.

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, and T. Florian Jaeger. 2015. Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science* 40:1382–1411.

Mateu, Jaume, and Víctor Acedo-Matellán. 2012. The manner/result complementarity revisited: A syntactic approach. In *The end of argument structure*, ed. María Cristina Cuervo and Yves Roberge, 209–228. Brill.

Melchin, Paul. 2019. The semantic basis for selectional restrictions. Doctoral Dissertation, University of Ottawa, Ottawa, ON.

References III

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. 1998. Building verb meanings. In *The projection of arguments: Lexical and compositional factors*, ed. Miriam Butt and Wilhelm Geuder, 97–134. Stanford, CA: CSLI.

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.

Reif, Emily, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of BERT. In *Advances in Neural Information Processing Systems*, ed. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, volume 32. Curran Associates, Inc. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf`.

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.

Sprouse, Jon, Carson T. Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. *Lingua* 134:219–248.

Wilcox, Ethan Gotlieb, Richard Futrell, and Roger Levy. 2024. Using computational models to test syntactic learnability. *Linguistic Inquiry* 55:805–848. URL `http://dx.doi.org/10.1162/ling_a_00491`.

Zehr, Jeremy, and Florian Schwarz. 2018. Penncontroller for internet based experiments (IBEX). Doi:10.17605/OSF.IO/MD832.