Understanding the role of linguistic distributional knowledge in cognition: A systematic comparison of tasks, models and parameters

Cai Wingfield
c.wingfield@lancaster.ac.uk

Louise Connell
l.connell@lancaster.ac.uk
@drlouiseconnell

Download PDF copy

Lancaster University

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 682848.

Linguistic distributional models (LDMs) learn associations between words from occurrences in language, and model how they can aid in cognitive tasks.

LDM research has recently focussed on certain neural-network models, optimising for similarity-driven tasks.

Do these recommendations generalise to modelling conceptual processing as a whole?

This study comprises a systematic comparison of LDMs on a wide range of tasks.

We ask: is there a one-size-fits-all LDM which works for general cognitive processing, and how performance varies by task features?

Using Bayesian model comparisons, we make recommendations as to the optimal LDM, corpus type and parameters for tasks with particular features.

We separately trained each LDM on each corpus with each parameter setting.

Each LDM was evaluated against each task dataset, with Bayesian comparison to a baseline model.

The models

N-gram
- Probability ratio n-gram
- Log n-gram
- PPMI n-gram

Count
- Probability ratio vector
- Log co-occurrence count vector
- Conditional probability vector
- PPMI vector

Predict
- CBOW
- Skip-gram

Context window
1 3 5 10

Corpus
BNC (100 M)
Subtitles (200 M)
UKWAC (2 B)

Vector distance
Euclidean
Cosine
Correlation

Embedding size
50 100 200 300 500

540 total

Task datasets
12 total

Concrete–abstract SDT
Semantic priming RT
Thematic association
Relatedness rating
Similarity rating
Synonym detection

Conceptual complexity

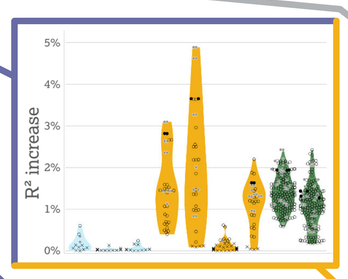
Explicit
- Response
- RT

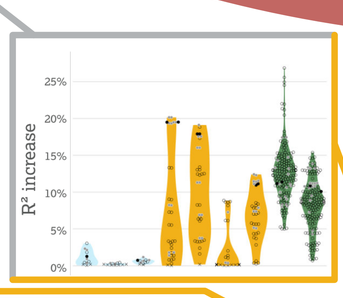
Implicit
- NT
- LDT
- J&M

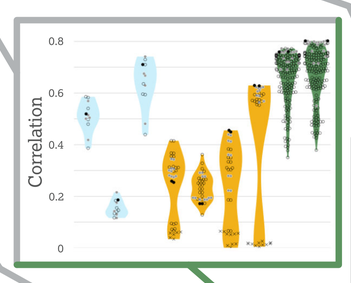
Explicit
- MEN
- Wordsim (sim) (rel)
- Simlex

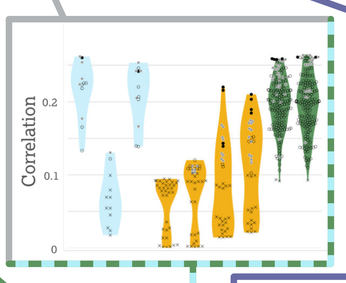
Measurement:
- LBM
- TOEFL
- ESL

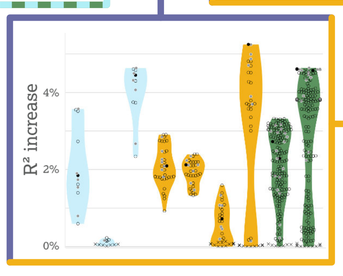
Results

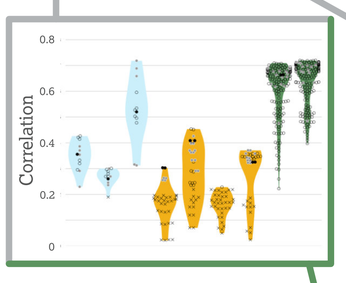


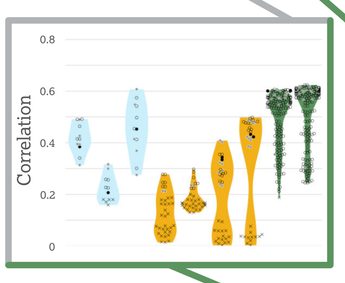


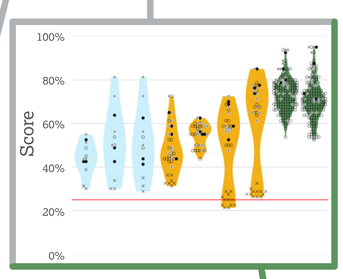


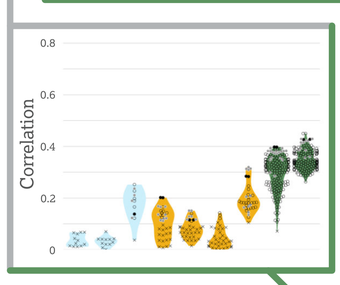


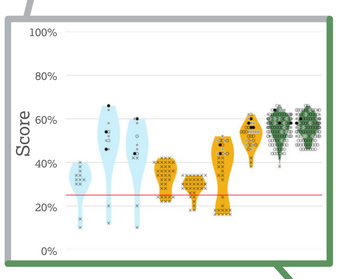


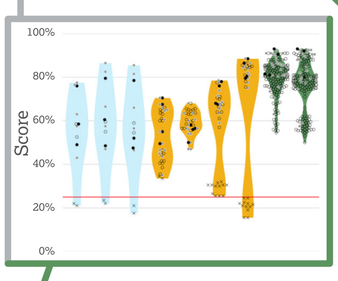












Optimal family:
- Predict
- Count

Optimal corpus:
- UKWAC
- BNC

Predict or n-gram
UKWAC or Subtitles

Count
Subtitles

N-gram models
- log n-gram
- Probability ratio n-gram
- PPMI n-gram

Count vector models
- log co-occurrence
- Conditional probability
- Probability ratio
- PPMI

Predict vector models
- Skip-gram
- CBOW

Findings

1 This is the widest systematic evaluation of LDMs on conceptual processing tasks used in cognitive psychology.

2 Corpus quality is often more important than corpus size.

3 Count models and large contexts excel at conceptually complex tasks and implicit measures.

4 Predict models and small contexts excel at explicit, similarity-based judgements.

5 Complexity and measurement variable are relevant to performance.

6 **No general-purpose LDM: optimal model depends on the properties of the task.**

References

BNC: BNC Consortium (2007) *The British National Corpus, version 3*.
ESL: Tatsuki (1998) Basic 2000 words-synonym match 1. *Interactive JavaScript Quizzes for ESL Students*.
J&M: Jouravlev & McRae (2016) Thematic relatedness production norms for 100 object concepts. *BRM* 48(4).
LBM: Levy, Bullinaria & McCormick (2017) Semantic vector evaluation and human performance on a new vocabulary MCQ test. *Proc. 39th Cog. Sci.*
MEN: Bruni, Tran & Baroni (2014) Multimodal distributional semantics. *J. AI Res.* 49(2014).
SDT: Pexman, Heard, et al. (2017). The Calgary semantic decision project: concrete/abstract decision data for 10,000 English words. *BRM* 49(2).
Simlex: Hill, Reichart & Korhonen (2016) Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Comp. Linguistics*.
SPP: Hutchison, Balota, et al. (2013) The semantic priming project. *BRM* 45(4).
Subtitles: van Heuven, Mandera, et al. (2014) Subtlex-UK: A new and improved word frequency database for British English. *QJ. Exp. Psychol* 67(6).
TOEFL: Educational Testing Service, Princeton NJ. (1989). Test of English as a foreign language.
UKWAC: Baroni, Bernardini, et al. (2009) The WaCky wide web: a collection of very large linguistically processed Web-crailed corpora. *Lang. Res & Eval.* 43(3).
Wordsim: Finkelstein, Gabrilovich, et al. (2001) Placing search in context: The concept revisited. *Proc. 10th WWW*.