Are longer verbal expressions really semantically more similar to each other? An investigation of the elaboration-bias in vector-based models of word meaning

Boris Forthmann – University of Münster
Fritz Günther – University of Milano-Bicocca
Rick Hass – Philadelphia University + Thomas Jefferson University
Mathias Benedek – University of Graz
Philipp Doebler – TU Dortmund University

psychoco 2020 – Dorthmund, 27th of February

Divergent thinking

„The unique feature of divergent production is that a variety of responses is produced“ (Guilford, 1959)

Divergent thinking

Is an indicator of…

- Everyday creative thinking ability (Kaufman & Beghetto, 2009)
- Creative potential (Lubart, Besançon, & Barbot, 2011; Runco & Acar, 2012)

Creative process (Mumford et al., 2008)

1. Problem definition
2. Information gathering
3. Concept selection
4. Conceptual combination
5. **Idea generation → divergent thinking**
6. Idea evaluation
7. Implementation
8. Monitoring

The Alternate Uses Task

- Instruction: *Please name as many different uses for a knife as possible.*

Idea	**Person**			
	1	2	3	4
as a wheapon	1	1	1	1
as a dart	0	1	1	0
as a screwdriver	1	0	1	0
as a cake server	0	0	0	1
stirring coffee	1	0	0	1

Reiter-Palmon, R., Forthmann, B., & Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 144-152.

Fluency Scoring

Idea	**Person**			
	1	2	3	4
as a wheapon	1	1	1	1
as a dart	0	1	1	0
as a screwdriver	1	0	1	0
as a cake server	0	0	0	1
stirring coffee	1	0	0	1
Fluency-Score	3	2	3	3

Uniqueness Scoring (Originality)

Idea	Person			
	1	2	3	4
as a wheapon	1	1	1	1
as a dart	0	1	1	0
as a screwdriver	1	0	1	0
as a cake server	0	0	0	1
stirring coffee	1	0	0	1
Uniqueness-Score	0	0	0	1
Uniqueness-Ratio	0	0	0	0.33

Forthmann, B., Paek, S. H., Dumas, D., Barbot, B., & Holling, H. (2019). Scrutinizing the basis of originality in divergent thinking tests: On the measurement precision of response propensity estimates. *British Journal of Educational Psychology*. Advance online publication.

Creative Quality Scores

- Originality (Wilson, Guilford, Christensen, 1953)
    - Uncommonness
    - Cleverness
    - Remoteness
- Appropriateness

Creative Quality Scores

- Originality (Wilson, Guilford, Christensen, 1953)
  - Uncommonness
  - Cleverness
  - Remoteness → semantic distance → vector-based models of word meaning
- Appropriateness

Vector-based models of word meaning – I

- All models represent word meanings as high-dimensional numerical vectors (i.e., semantic space)
- These models allow computing of the semantic similarity between any pair of words (or larger expressions) as cosine similarity between their respective vectors
- These models predict a variety of human behavior:
  - Categorization tasks
  - Synonym tests
  - Similarity judgments
  - Lexical priming

Vector-based models of word meaning – II

- Latent Semantic Analysis (LSA; Landauer & Dumais, 1997)
  - Word-by-document co-occurrences
  - Weighting schemes (e.g., pointwise mutual information)
  - Dimensionality reduction (e.g., singular value decomposition)
- Hyperspace Analogue to Language model (HAL; Lund & Burgess, 1996)
  - Based on word-by-word co-occurrences
  - Weighting schemes and dimensionality reduction (analogous to LSA)
- Continuous Bag of Words model (CBOW as part of word2vec; see Mikolov et al., 2013)

Vector-based models of word meaning – III

- Continuous Bag of Words model (CBOW as part of word2vec; see Mikolov et al., 2013)
  - Based on a neural network architecture
  - Target words are predicted by sorrounding words

Why Using Vector-based Models of Word Meaning?

1. Scoring is objective
2. The models are empirically validated
3. The models are theoretically justified
4. Scoring is less labor intensive as compared to other scorings
5. There are freely available tools to apply the models

Study 1 – Forthmann et al. (2017)

Participants: *N* = 199 (female = 142; age: *M* = 24.48, *SD* = 6.86)

DT tasks: Alternate Uses (rope, garbage bag, paperclip); 2.5 minutes; be-creative instructions

Scoring:

Overall quality (Ratings)

Cleverness (Ratings)

Uncommonness (Statistical Frequency)

Semantic Distance (LSA)

Complexity/Elaboration (number of characters)

Forthmann, B., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2017). Typing speed as a confounding variable and the measurement of quality in divergent thinking. *Creativity Research Journal*, *29*(3), 257-269.

Results – Study 1 – Forthmann et al. (2017)

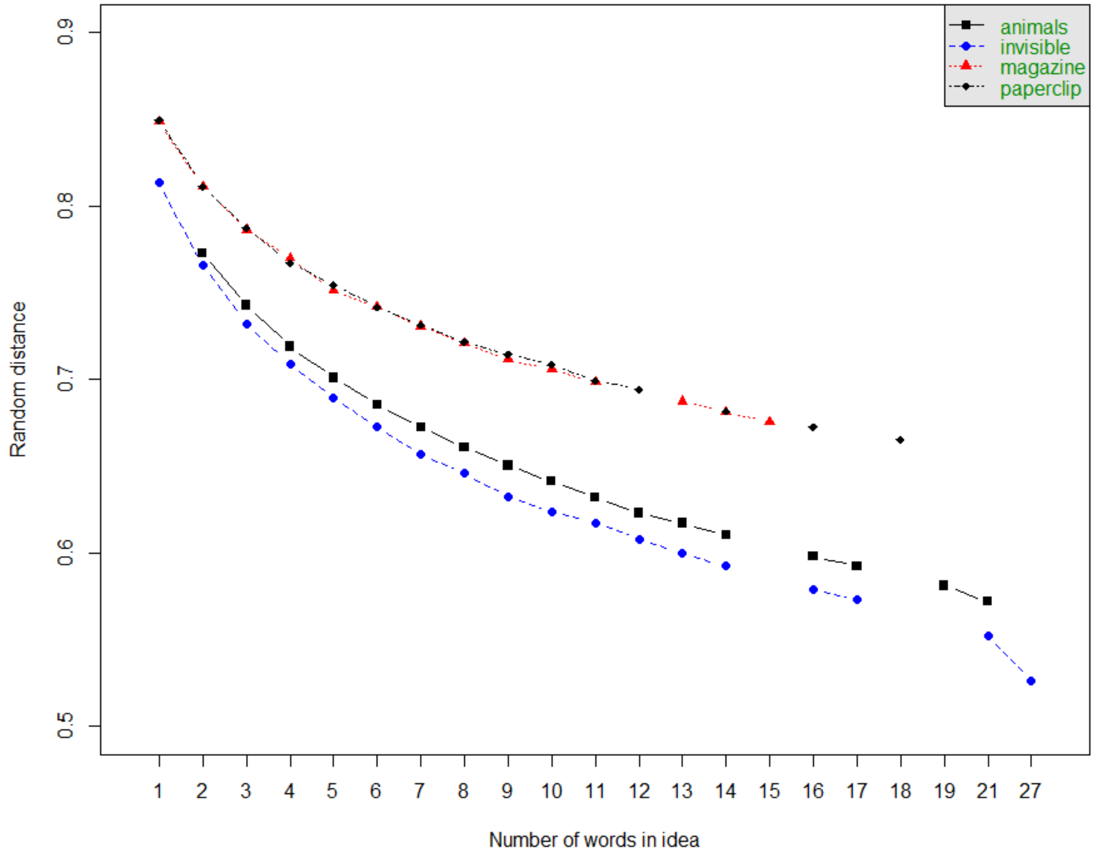
TABLE 3
Correlation matrix for the latent variables in Study 2 – with and without statistical control of confounding variables

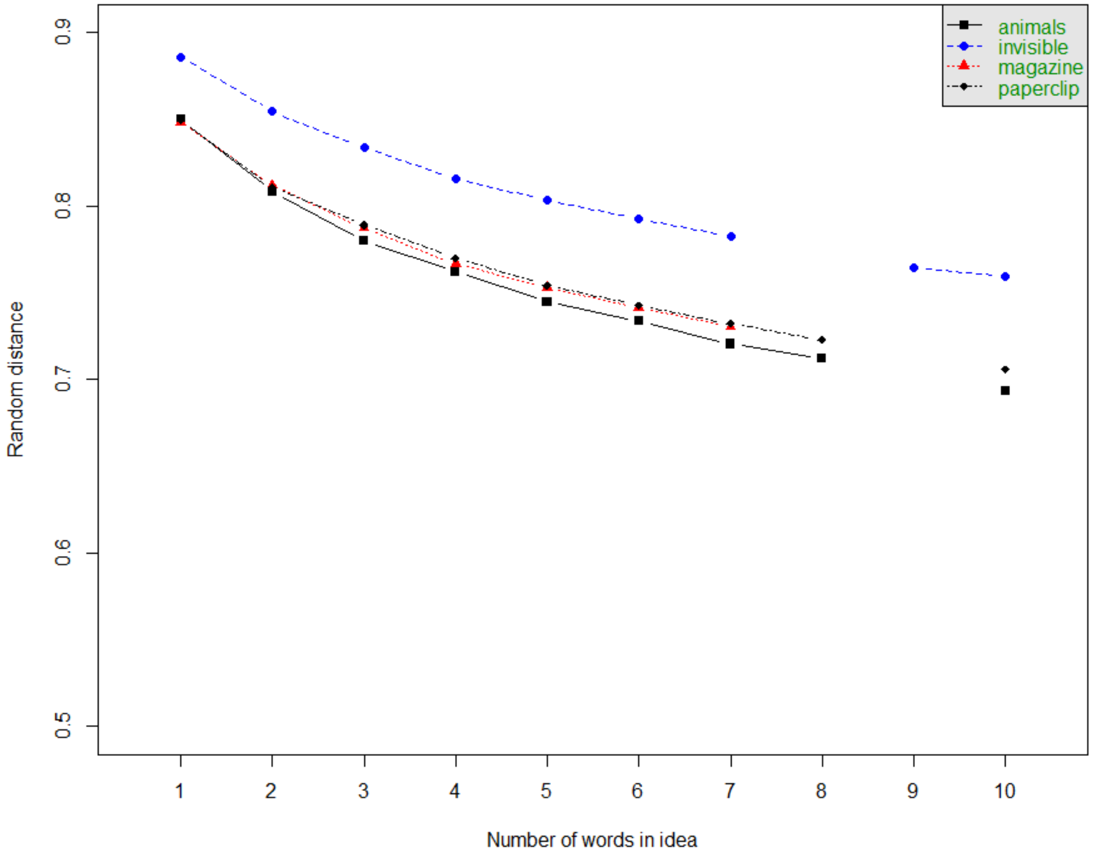
		1	*2*	*3*	*4*	*5*	*6*
Overall quality	1	–	.822***	.589***	.553***	.419***	.017
Cleverness	2	.831***	–	.454**	.240	.462***	.206
Uncommonness	3	.452***	.503***	–	.444*	.142	NA
Remoteness	4	−.005	−.215*	.307**	–	NA	.052
Complexity	5	.437***	.481***	−.091	−.793***	–	−.324***
Fluency	6	.061	.253*	.737***	.288**	−.283***	–
Typing speed	7	.217*	.290***	.261**	.012	.151*	.205**

Notes. $n = 199$. Lower triangle: Correlations of the latent variables as they were measured as depicted in Figure 1. Upper triangle: Correlations of the latent variables after confounding variables are statistically controlled. Typing speed was controlled in all variables. Fluency was additionally controlled in uncommonness. Complexity was additionally controlled in remoteness. NA = not applicable.
$*p < .05$. $**p < .01$. $***p < .001$.

Forthmann, B., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2017). Typing speed as a confounding variable and the measurement of quality in divergent thinking. *Creativity Research Journal*, *29*(3), 257-269.

Study 2 – Simulation Results (LSA semantic distance) – Forthmann et al. (2019)



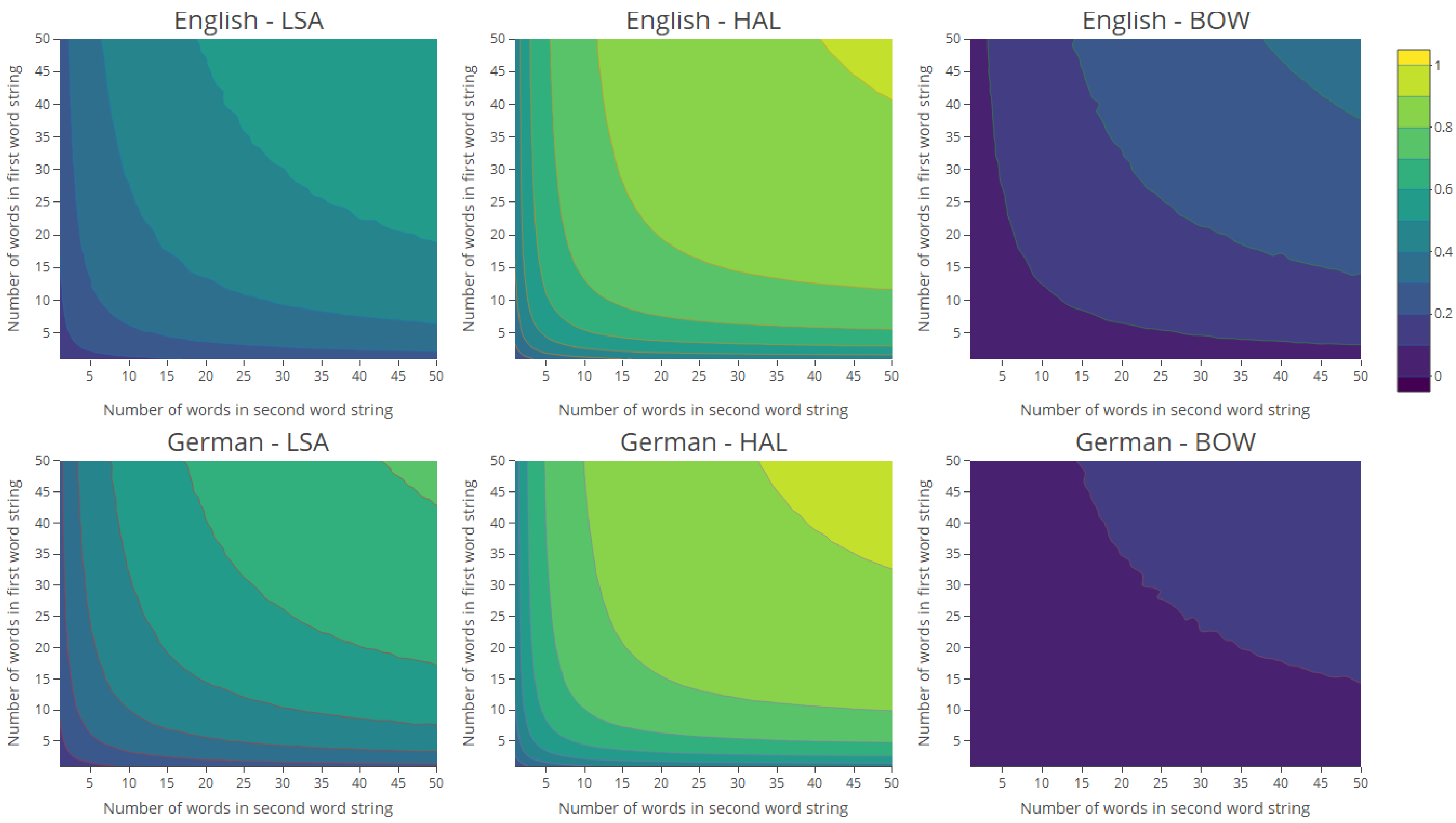


Forthmann, B., Oyebade, O., Ojo, A., Günther, F., & Holling, H. (2019). Application of latent semantic analysis to divergent thinking is biased by elaboration. *The Journal of Creative Behavior*, *53*(4), 559-575.

Open Questions

- Does the elaboration bias generalize to other vector-based models of word meaning?
- How does the bias emerge?
- Are computationally less intensive bias-corrections available as compared to a simulation-based correction (Forthmann et al., 2019)?

Generalization check

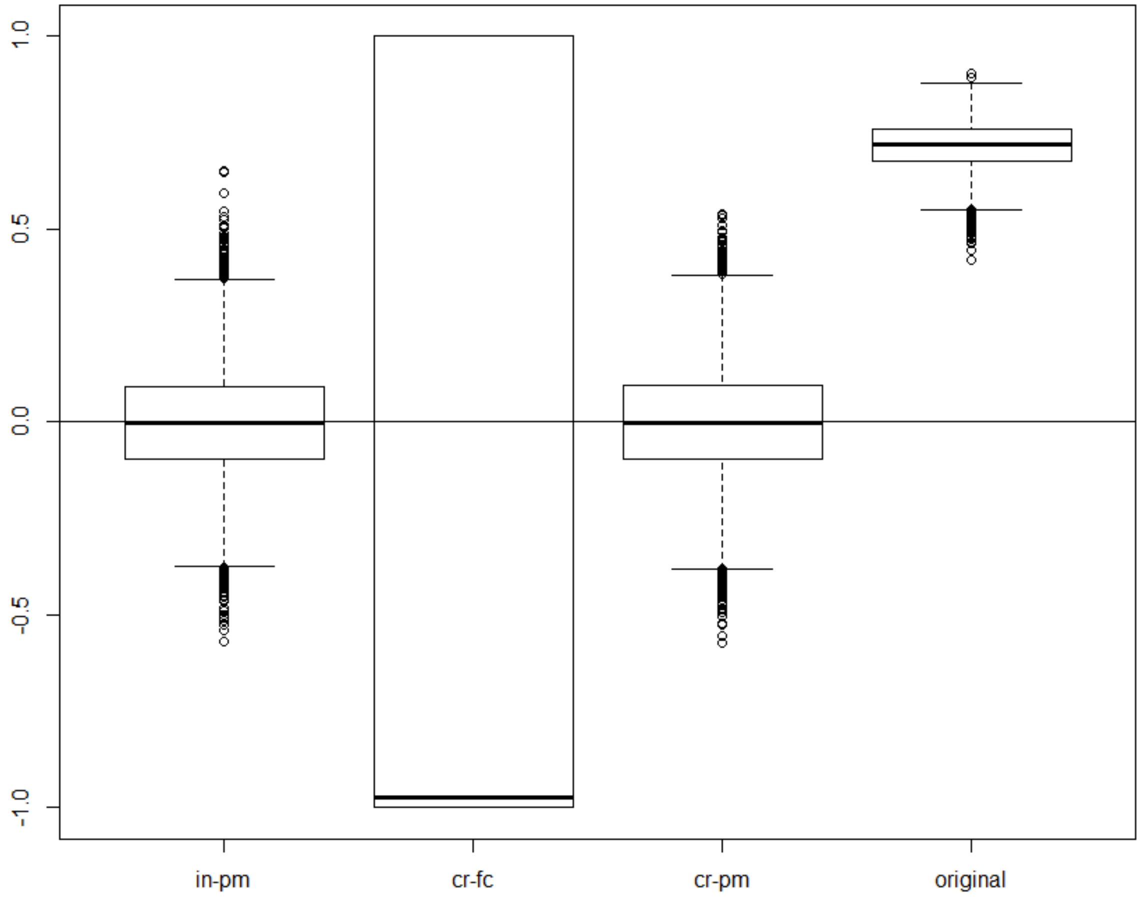


How does the bias emerge?

- The bias occurs when at least one of the column means of the semantic space is different from zero

How can we mitigate the bias without simulations?

- Centering of ranked columns
- Inverse normal transformation of the columns
- Transformations applied only to the first component
- Transformations combined with postmultiplication of the column standard deviations



Do these transformations work?

- For English spaces 9 benchmarks were checked (1 synonym, 5 rating, 3 categorization)
- For German spaces 3 benchmarks were checked (2 rating, 1 categorization)

→ In 8 cases out of the 12 benchmark checks HAL with inverse normal transformation yielded the best performance

Questions? Discussion points?

Thank you for your interest!