# Language models can outperform empirical predictability in predicting eye movement data

**Poster** · August 2019

DOI: 10.13140/RG.2.2.27252.50562



**4 authors:**

Markus J. Hofmann
Bergische Universität Wuppertal


Steffen Remus
University of Hamburg


Chris Biemann
University of Hamburg


Ralph Radach
Bergische Universität Wuppertal

# Language models can outperform empirical predictability in predicting eye movement data


Markus J. Hofmann[1]*, Steffen Remus[2], Chris Biemann[2], & Ralph Radach[1]

[1] General and Biological Psychology, University of Wuppertal [2] Language Technology, University of Hamburg * Email: mhofmann@uni-wuppertal.de


**Introduction: Language Models provide a deep explanation of semantic long-term memory, because they reflect the 3 basic stages of the memory process…**

(e.g., Hofmann et al., 2018)

## Experience:

### Training Corpora

*English:*
**WIKI:** Wikipedia dump, 5M documents/articles, 87M sentences, 1.8B words
-> general knowledge
**SUB**: Subtitles from opensubtitles.org, 110K documents/movies, 128M sentences/utterances, 716M words
-> spoken language.
-> SUB-frequency correlates better with lexical decision/naming times (e.g. Brysbaert et al., 2011)

*German*:
**WIKI**: 114K articles / 7.7M sentences / 180M words
**SUB**: 7K movies / 7.3M utterances / 54M words

## Consolidation: Language Models (LMs)

### N-gram

Conditional probability that a word N occurs, given the two previous words, e.g.

N-2 N-1 N
*We enjoyed the tasty ice cream.*

-> *P(N | N-1, N-2)*

McDonald & Schillcock (2003): Transitional probability ~ 2-gram predicts SFD & GD

Wang et al. (2010): TVT, but not first fixation effects of 2-gram models in Chinese

Smith & Levy (2013): Kneser & Ney (1995) smoothed 3-gram model predicted GD data

-> Early and/or late proceses

### Topic

Probability of **w**ords to occur together in **d**ocuments is approximated by N = 200 latent **topics z**

$$p(w|d) = \sum_{i=1..N} p(w|z_i)p(z_i|d)$$

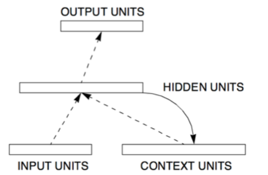
-> Word-topic- & topic-document-matrices

Document-based like Latent Semantic Analysis (LSA, Landauer & Dumais, 1997)
At retrieval: d ~ preceding words in sentence:

Pynte et al. (2008) found LSA-effects in GD and Wang et al. (2010) in TVT
-> Late processes?
Griffiths et al (2007): Topics often better than LSA

-> Can topical matches to previous sentence context account for **long-range semantics?**

### Recurrent Neural Network (RNN)




Input : Present word
Output: Next word to be predicted
Hidden Units: "Semantic" representation
Context: Hidden units of last word copied:
Trainable Connections:
Initialized with random weights
Training: The lower the output probability of the next word, the larger is the connection weight change

Frank & Brod (2011): RNN predicts GD

Elman (1990); Mikolov (2012)

## Retrieval:

### Eye Movement predictions

Eye movements (**EM**) and incremental Cloze Completion Probability (**CCP**-) based predictability data from Engelmann et al. (2013):
- **SRC**: Schilling, Rayner & Chumbley (1998)
- **PSC**: Potsdam Sentence Corpus (Kliegl et al., 2006)

Excluding first and last words of the sentences

Outliers:

Single-fixation duration (**SFD**, 70-800ms)

First-pass gaze duration (**GD**, 70-1200ms),

Total Viewing Time (**TVT**, 70-1600ms

| | Sentences | Targets | Participants | | Rows of data in EM analysis | | |
|---|---|---|---|---|---|---|---|
| | | | CCP | EM | SFD | GD | TVT |
| SRC | 48 | 536 | 20 | 30 | 6,451 | 8,671 | 8,736 |
| PSC | 144 | 850 | 272 | 222 | 100,975 | 134,835 | 135,021 |

**Table 1:** Sample sizes of EM and CCP data

Comparing CCP-based to LM-based "predictability": Present word, **last** word: Lag effects, **next** word: Successor effects

## Statistical Analyses

***Generalized Additive Models (GAM):*** Thin plate regression splines (mgcv-package in R); Gamma functions with a logarithmic link (cf. Smith & Levy, 2013)

*Stepwise GAMs:*

Baseline predictors: Random terms: **sub**ject and **items** (M01); **Land**ing site (M02); **Length and freq**uency of the present (M03), **last** (M04), and **next** word (M05).

**Different types of "predictability":**

Cloze-Completion-Based (**CCP**) predictability scores of present (M06), last (M07) and next word (M08):

**WIKI**pedia-trained (M09-M17) & **SUB**titles-trained (M15-M27): N-gram (M09-M11 & M18-M20); Topic (M12-M14 & M21-M23); RNN (M15-M17 & M24-M26)

All **WIKI**-trained LMs better than **CCP**? (M27)

**SUB** better than **WIKI**? (M28)

Both better than SUB? (M29)

Does **CCP** account **for additional variance**? (M30)

*Model scores*: Generalized Cross-Validation (**GCV**) scores: Mean prediction error based on a leave-one-out cross validation process (~ AIC/BIC; lower ~ better); **%ΔR²**; log likelihood tests whether present GAM fits better than last: **Deviance:** Similar to root-mean-squared error. Positive value indicates the present GAM is better; Positive **df** indicates present GAM is simpler; * P< 0.05.

## Stepwise GAMs

| | | SRC | | | | | | | | | PSC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | SFD | | | GD | | | TVT | | | SFD | GD | TVT |
| Model + Predictors | | GCV | %ΔR² | Deviance (df) | GCV | %ΔR² | Deviance (df) | GCV | %ΔR² | Deviance (df) | %ΔR² | %ΔR² | %ΔR² |
| M01 | Random: item + sub | 0.1324 | 0.38 | M01 | 0.1803 | 0.4 | M01 | 0.2154 | 0.26 | M01 | 0.05* | 0.13* | 0.09* |
| M02 | M01 + land | 0.1318 | 0.52 | 5.9(8.1)* | 0.1777 | 1.32 | 25(8.6)* | 0.2109 | 1.9 | 42.8(8.8)* | 0.21* | 1.63* | 1.75* |
| M03 | M02 + len + freq | 0.1283 | 2.44 | 22.7(4.1)* | 0.1675 | 5.39 | 90.7(9.1)* | 0.1969 | 5.97 | 124.8(9)* | 1.67* | 7.77* | 6.88* |
| M04 | M03 + len_last + freq_last | 0.1275 | 0.6 | 7.2(6.8)* | 0.166 | 0.75 | 16.3(13.2)* | 0.1942 | 1.15 | 27.3(12.4)* | 1.8* | 0.85* | 0.88* |
| M05 | M04 + len_next + freq_next | 0.1271 | 0.41 | 5.8(14.2)* | 0.1653 | 0.51 | 13.9(22.9)* | 0.1932 | 0.48 | 14.7(16.3)* | 0.34* | 0.33* | 0.48* |
| ***CCP*** | | | | M05 | | | M05 | | | M05 | M05 | | |
| M06 | M05 + ccp | 0.1271 | 0.02 | 0.4(1.8) | 0.1653 | 0.03 | 0.8(2.9) | 0.1929 | 0.27 | 6.4(8.9)* | 0.09* | 0.09* | 0.3* |
| M07 | M06 + ccp_last | 0.127 | 0.04 | 1(2.9) | 0.1651 | 0.13 | 4.6(9.4)* | 0.193 | -0.13 | -3.1(-4.7)* | 0.05* | 0.05* | 0.13* |
| M08 | M07 + ccp_next | 0.1269 | 0.02 | 0.5(1.1) | 0.1648 | 0.08 | 2.1(0.5)* | 0.1929 | 0.01 | 0.9(1.4) | 0.32* | 0.18* | 0.09* |
| ***Wikipedia, N-gram*** | | | | M05 | | | M05 | | | M05 | M05 | | |
| M09 | M05 + ngram | 0.1271 | 0 | 0.1(1.1) | 0.1651 | 0.12 | 2.5(5.4)* | 0.1927 | 0.3 | 6(5.5)* | 0.29* | 0.14* | 0.27* |
| M10 | M09 + ngram_last | 0.1271 | 0.02 | 0.9(6.3) | 0.1652 | -0.1 | -2.9(-5.3)* | 0.1928 | -0.01 | 0.1(1.1) | 0.12* | 0.08* | 0.13* |
| M11 | M10 + ngram_next | 0.1272 | -0.08 | -1(-3.8) | 0.1652 | 0.09 | 2.9(7.2)* | 0.1928 | 0 | 0.6(2.1) | 0.11* | 0.09* | 0.06* |
| ***Wikipedia, Topic*** | | | | M05 | | | M05 | | | M05 | M05 | | |
| M12 | M05 + topic | 0.127 | 0.12 | 2.2(8.3)* | 0.1652 | 0.06 | 2.6(8.5) | 0.1932 | 0.03 | 0.8(1.4) | 0.32* | 0.19* | 0.18* |
| M13 | M12 + topic_last | 0.1269 | 0.1 | 1.9(7) | 0.165 | 0.18 | 4.1(7.2)* | 0.1929 | 0.2 | 5.7(9.6)* | 0.2* | 0.15* | 0.14* |
| M14 | M13 + topic_next | 0.1268 | 0.06 | 0.7(1.4)* | 0.1647 | 0.12 | 1.2(-3.9) | 0.1927 | 0.15 | 6.6(14.6)* | 0.05* | 0.02* | 0.03* |
| ***Wikipedia, RNN*** | | | | M05 | | | M05 | | | M05 | M05 | | |
| M15 | M05 + rnn | 0.127 | 0.03 | 0.7(3.8) | 0.1652 | 0.09 | 2.3(6.6) | 0.1929 | 0.2 | 5.1(6.6)* | 0.19* | 0.17* | 0.28* |
| M16 | M15 + rnn_last | 0.127 | 0.03 | 0.6(2.5) | 0.1653 | -0.09 | -2.6(-5.2)* | 0.1929 | 0.01 | 1.1(4) | 0.14* | 0.1* | 0.13* |
| M17 | M16 + rnn_next | 0.1269 | 0.14 | 1.9(4.4)* | 0.1649 | 0.26 | 7.5(15.2)* | 0.1924 | 0.24 | 6.9(8.6)* | 0.08* | 0.12* | 0.09* |
| ***Subtitles, N-gram*** | | | | M05 | | | M05 | | | M05 | M05 | | |
| M18 | M05 + ngram | 0.1268 | 0.27 | 2.9(5.5)* | 0.1648 | 0.31 | 3.9(0.6)* | 0.1925 | 0.38 | 8.4(7.3)* | 0.26* | 0.23* | 0.31* |
| M19 | M18 + ngram_last | 0.1263 | 0.41 | 5.1(8.4)* | 0.1645 | 0.14 | 3.8(5.4)* | 0.1924 | 0.08 | 2.6(5.5) | 0.13* | 0.12* | 0.22* |
| M20 | M19 + ngram_next | 0.1263 | -0.03 | -0.4(-0.9) | 0.1645 | 0.03 | 1.3(6) | 0.1924 | 0.02 | 1.9(6.1) | 0.08* | 0.04* | 0.06* |
| ***Subtitles, Topic*** | | | | M05 | | | M05 | | | M05 | M05 | | |
| M21 | M05 + topic | 0.127 | 0.13 | 1.6(5.2)* | 0.1652 | 0.06 | 1.3(4.1) | 0.1931 | 0.16 | 4.4(8.1)* | 0.2* | 0.15* | 0.13* |
| M22 | M21 + topic_last | 0.1269 | 0.03 | 0.5(1.7) | 0.1653 | -0.01 | 0.1(3) | 0.1931 | -0.01 | 1.9(8) | 0.27* | 0.14* | 0.14* |
| M23 | M22 + topic_next | 0.1269 | 0.01 | 0.3(1.7) | 0.1653 | -0.01 | 0.1(1.1) | 0.1932 | -0.01 | 0(0.9) | 0.07* | 0.02* | 0.02* |
| ***Subtitles, RNN*** | | | | M05 | | | M05 | | | M05 | M05 | | |
| M24 | M05 + rnn | 0.1271 | -0.07 | -0.9(-3.2) | 0.1653 | 0 | 0.1(0.9) | 0.1932 | 0.12 | 3.5(7.8)* | 0.25* | 0.21* | 0.32* |
| M25 | M24 + rnn_last | 0.127 | 0.07 | 0.8(0.3)* | 0.1652 | 0.17 | 3.6(8.9)* | 0.1933 | -0.14 | -3.4(-6.2)* | 0.03* | 0.02* | 0.1* |
| M26 | M25 + rnn_next | 0.1269 | 0.07 | 0.6(1.1)* | 0.165 | 0.1 | 1.6(0.7)* | 0.1931 | 0.1 | 2(0.7)* | 0.16* | 0.12* | 0.04* |
| ***N-gram, Topic, RNN vs. CCP*** | | | | M08 | | | M08 | | | M08 | M08 | | |
| M27 | WIKI: M11+ M14 + M17 | 0.1268 | 0.35 | 5.7(22.5)* | 0.1642 | 0.5 | 10.3(20.4)* | 0.1917 | 0.82 | 24.4(40.1)* | 0.58* | 0.41* | 0.42* |
| M28 | SUB: M20 + M23 + M26 | 0.1261 | 0.61 | 6.2(6.4)* | 0.1641 | 0.09 | 0.7(-0.7) | 0.1923 | -0.41 | -11.8(-18.3)* | 0.08 | 0.13* | 0.08* |
| M29 | WIKI + SUB: M27 + M28 | 0.1259 | 0.34 | 5.5(21)* | 0.1633 | 0.74 | 19.5(44.5)* | 0.1912 | 0.73 | 23.7(44.3)* | 0.57* | 0.46* | 0.73* |
| ***CCP, N-gram, Topic, RNN*** | | | | M29 | | | M29 | | | M29 | M29 | | |
| M30 | M08 + M29 | 0.126 | -0.05 | -0.8(-3.9) | 0.1633 | 0.08 | 2.7(12.2) | 0.1912 | -0.01 | -1.5(-2.9) | 0.2* | 0.18* | 0.3* |

## M30: F-values

| | SRC | | | PSC | | |
|---|---|---|---|---|---|---|
| | SFD | GD | TVT | SFD | GD | TVT |
| sub | 7.768** | 13.607*** | 0 | 59.085*** | 196.011*** | 94.654*** |
| item | 5.335* | 4.805* | 5.57* | 0.163 | 3.566* | 2.792+ |
| land | 2.744+ | 8.662*** | 10.589*** | 11.87*** | 15.445*** | 10.397*** |
| len | 0.17 | 9.254*** | 16.643*** | 21.343*** | 337.608*** | 242.906*** |
| fc | 6.226*** | 3.914*** | 4.419*** | 26.857*** | 53.876*** | 60.133*** |
| len_last | 9.704*** | 8.388*** | 15.369*** | 24.403*** | 31.613*** | 28.384*** |
| fc_last | 0.684 | 0.916 | 2.893* | 8.898*** | 9.141*** | 11.75*** |
| len_next | 4.636* | 1.978+ | 3.111*** | 11.814*** | 13.386*** | 29.691*** |
| fc_next | 3.309*** | 3.194** | 3.949*** | 13.737*** | 24.523*** | 16.909*** |
| ngram_wiki | 0.004 | 1.47 | 0.931 | 8.103*** | 4.803*** | 5.547*** |
| ngram_wiki_last | 0.529 | 1.309 | 1 | 2.312* | 3.074** | 5.672*** |
| ngram_wiki_next | 0.322 | 2.864+ | 2.611 | 10.545*** | 10.299*** | 17.936*** |
| topic_wiki | 3.522+ | 2.01+ | 2.053 | 17.229*** | 16.185*** | 15.552*** |
| topic_wiki_last | 0.076 | 2.868** | 1.821 | 3.804*** | 8.11*** | 11.517*** |
| topic_wiki_next | 1.955 | 2.841** | 2.993+ | 2.124* | 2.203* | 3.902*** |
| rnn_wiki | 0.763 | 1.03 | 2.064* | 2.908** | 6.026*** | 10.897*** |
| rnn_wiki_last | 2.169 | 0.37 | 0.989 | 4.778*** | 9.175*** | 11.3*** |
| rnn_wiki_next | 3.348** | 3.046** | 3.68*** | 5.835*** | 7.982*** | 15.799*** |
| ngram_sub | 2.889** | 5.683*** | 4.71** | 8.349*** | 15.967*** | 15.511*** |
| ngram_sub_last | 3.824*** | 3.85** | 2.279* | 7.65*** | 8.307*** | 12.591*** |
| ngram_sub_next | 1.152 | 1.636 | 1.333 | 6.878*** | 14.211*** | 19.518*** |
| topic_sub | 2.443* | 0.427 | 2.701* | 14.601*** | 15.87*** | 21.458*** |
| topic_sub_last | 1.067 | 1.725+ | 2.289 | 12.216*** | 12.563*** | 11.359*** |
| topic_sub_next | 0.212 | 0.02 | 0.452 | 5.555*** | 6.839*** | 14.955*** |
| rnn_sub | 0.184 | 1.747 | 0.029 | 4.291*** | 3.345** | 2.833** |
| rnn_sub_last | 3.8+ | 2.083* | 0.498 | 7.242*** | 5.918*** | 8.929*** |
| rnn_sub_next | 2.096* | 1.327 | 0.807 | 7.485*** | 8.175*** | 4.527*** |
| pred | 1.108 | 0.025 | 1.084 | 11.417*** | 18.473*** | 34.172*** |
| pred_last | 1.923+ | 1.315 | 1.038 | 2.233+ | 7.096*** | 16.105*** |
| pred_next | 0.547 | 1.667 | 1.526 | 15.085*** | 9.84*** | 7.767*** |
| Total %R² (adj.) | 5.69 | 10.0 | 11.0 | 5.96 | 12.2 | 12.21 |

**Table 2:** Significantly better predictions than previous model in blue, worse predictions in red (SRC). In PSC, all GAMs provided better predictions than the previous GAMs. Therefore, we only report (relatively large) increases in explained variance (in blue, if SRC provided significant increases in the same comparison) (*P < 0.05).

**Table 3:** Significant predictors in M30. (+ P < .1; * P < 0.05; ** P < 0.01; *** P < 0.001)

## Single LMs vs. CCP

| | SRC | | | PSC | | |
|---|---|---|---|---|---|---|
| | SFD | GD | TVT | SFD | GD | TVT |
| N-gram: WIKI (M11) | -1.9(-2.3)* | -5.1(-5.6)* | 8.8(20.1)* | 5.2(-1.6) | 4.2(-0.5) | -8.7(0)* |
| Topics: WIKI (M14) | 3(10.9)* | 0.3(-1.1) | 8.9(13.6)* | 11.6(0.2)* | 8.3(-0.8) | -70.8(-0.6)* |
| RNN: WIKI (M17) | 1.4(4.8) | -0.3(3.7) | 8.7(13.2)* | -6.8(0) | 9.8(-0.7) | -21.9(-0.5)* |
| N-gram: SUB (M20) | 5.7(7.1)* | 1.5(-1) | 2.1(11.4) | 2.2(-0.3) | 21.7(-0.4) | 34.4(-0.2) |
| Topics: SUB (M23) | 0.5(2.8) | -6.1(-4.7)* | -2.1(-3.3)* | 10.6(0.2)* | 3.7(-1.2) | -82.8(0.1) |
| RNN: SUB (M26) | -1.4(-7.7) | -2.3(-2.4)* | 8.8(20.1)* | -3.3(0.4) | 3.5(-2.3) | -20.5(-1.1)* |

**Table 4**: Deviance (df) and * (P< 0.05). Blue = LM is better; red = CCP is better

## Take-Home:

- 3 WIKI-based LMs together provide better predictions than CCP (M27; **Table 2**)
- No general training corpus advantage, SUB better in 3, Wiki in 1 out of 6 comparisons (M28)
- Taking together both training corpora improves the predictions: general knowledge + spoken language (M29)
- > Replicating and extending Hofmann et al.'s (2017) linear approaches to SFD(, N400 and CCP) data for the PSC
- When LMs are trained by larger English corpora, a small CCP sample doesn't improve EM prediction (SRC). In a large CCP sample and smaller German training corpora, CCP improves EM predictions (PSC; M30)
- Single LMs vs. CCP: LMs win in SRC (6 vs. 5), but loose in PSC (2 vs. 4, smaller LM training corpora, larger CCP-sample). CCP is better for GD data, but LMs are often better for predicting SFD data (**Table 4**)
- M30: In a large EM data set (PSC), all predictors except some item random terms are significant predictors. Nearly all "item-level" variance explained? (**Table 3**)
- **Topic**: Lower F-values than N-gram & RNN in SRC, but highest in PSC for TVT: Late long-range semantics? Effects are distributed across training corpora, EM measures and position (WIKI_last+next~GD, SUB~SFD,TVT)
- **RNN**-based predictability of next word is best in WIKI training: general(ized) knowledge?
- **N-gram**-based conditional probability of present and last word is best in SUB training: spoken language?

## References


Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). A review of recent developments and implications for the choice of frequency estimates in German. *Exp. Psychol.*, *58*, 412–424.

Elman, J. (1990). Finding structure in time. *Cog. Sci.*, *14*, 179– 211.

Engelmann, F., Vasishth, S., Engbert, R., & Kliegl, R. (2013). A framework for modeling the interaction of syntactic processing and eye movement control. *Topics in Cogn. Sci.*, *5*(3), 452–474.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychol. Sci.*, *22*(6), 829–834.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychol. Rev.*, *114*(2), 211–244.

Hofmann, M. J., Biemann, C., & Remus, S. (2017). Benchmarking n-grams, topic models and recurrent neural networks by cloze completions, EEGs and eye movements. In B. Sharp, F. Sedes, & W. Lubaszewsk (Eds.), *Cognitive Approach to Natural Language Processing* (pp. 197–215). London, UK: ISTE Press.

Hofmann, M. J., Biemann, C., Westbury, C. F., Murusidze, M., Conrad, M., & Jacobs, A. M. (2018). Simple co-occurrence statistics reproducibly predict association ratings. *Cogn. Sci.*, *42*, 2287–2312.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the Mind During Reading : The influence of past, present, and future words on fixation durations. *J. Exp. Psychol.: Gen.*, *135*(1), 12–35.

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 181–184). Detroit, MI, USA: IEEE.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.*, *104*(2), 211–240.

McDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vis. Res.*, *43*(16), 1735–1751.

Mikolov, T. (2012). *Statistical language models based on neural networks (PhD thesis)*. Brno university of technology, Brno, Czech Republic.

Pynte, J., New, B., & Kennedy, A. (2008). On-line contextual influences during reading normal text: A multiple-regression analysis. *Vis. Res.*, *48*(21), 2172– 2183.

Schilling, H. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: word frequency effects and individual differences. *Mem. & Cogn.*, *26*(6), 1270–1281.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Wang, H. C., Pomplun, M., Chen, M., Ko, H., & Rayner, K. (2010). Estimating the effect of word predictability on eye movements in Chinese reading using latent semantic analysis and transitional probability. *Quart. J. Exp. Psychol.*, *63*(7), 1374-1386.