# Predict but Also Integrate: an Analysis of Sentence Processing Models for English and Hindi

**Nina Delcaro**
Eindhoven University of Technology,
The Netherlands
n.delcaro@student.tue.nl

**Luca Onnis**
Linguistics and Scandinavian Studies,
University of Oslo, Norway
lucao@uio.no

**Raquel G. Alhama**
Institute for Logic, Language and Computation
University of Amsterdam,
The Netherlands
rgalhama@uva.nl

## Abstract

Fluent speakers make implicit predictions about forthcoming linguistic items while processing sentences, possibly to increase efficiency in real-time comprehension. However, the extent to which prediction is the primary mode of processing human language is widely debated. The human language processor may also gain efficiency by integrating new linguistic information with prior knowledge and the preceding context, without actively predicting. At present, the role of probabilistic integration, as well as its computational foundation, remains relatively understudied. Here, we explored whether a Delayed Recurrent Neural Network (d-RNN, Turek et al., 2020), as an implementation of both prediction and integration, can explain patterns of human language processing over and above the contribution of a purely predictive RNN model. We found that incorporating integration contributes to explaining variability in eye-tracking data for English and Hindi.

## 1 Introduction

Languages are acquired and processed in real time. The transient quality of spoken language is evident, as it vanishes the moment it is spoken. And while written words appear fixed on a page, skilled readers assimilate them in sequence rapidly, seldom needing to double-back and review past words, and even skipping words entirely. This transitory aspect of language, coupled with the remarkable efficiency and speed at which humans use it, suggests that our brains harness specialized processes for managing information that fluidly unfolds in a sequence.

One proposed cognitive mechanism is prediction, the process by which a listener or reader anticipates upcoming linguistic information during language comprehension. This anticipation is based on internalized knowledge of language, previous local context information, and accumulated world-knowledge from semantic and episodic long-term memory. Psycholinguistic research suggests that individuals often implicitly predict elements such as the next word or grammatical structure while engaging with language, which allows for more efficient processing and understanding (Hale, 2016; Pimentel et al., 2023; Wilcox et al., 2023a). Prediction can occur at multiple levels, from anticipating the completion of a familiar phrase to forecasting the thematic content of a conversation or narrative. Probabilistic word prediction has been explicitly implemented in a class of cognitive recurrent models since the inception of the Recurrent Neural Network (Elman, 1990).

However, an unresolved debate in psycholinguistics centers around the extent to which the human language processor anticipates upcoming information (prediction) and how it assimilates incoming linguistic information with existing knowledge (integration, Ferreira and Chantavarin, 2018; Kuperberg and Jaeger, 2016; Nieuwland et al., 2020). A mechanism of probabilistic integration would not necessarily try to predict upcoming material, but instead increase efficiency by evaluating the probability of the preceding context given each heard/read linguistic item (e.g., the current word). In recent work (Onnis and Huettig, 2021; Onnis et al., 2022) this mechanism has been modeled successfully using n-gram language models as the backward transitional probability, *P(prior context | word)* as a proxy for integration, as opposed to prediction in the form of forward transitional probability, *P(word | current word)*.

Here, we conducted an exploratory analysis to determine whether a recurrent language model implementing integration can explain patterns of human language processing (online reading times revealed by eye movements from existing psycholinguistic datasets) over and above the contribution of a purely predictive language model. We did so by comparing two types of Recurrent Neural Networks (RNNs), namely the classic RNN, and the Delayed



*Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 101–108

RNN (d-RNN) as proposed by Turek et al. (2020). The classic RNN can be considered an implementation of prediction, while the d-RNN implements both prediction and integration (see details below). To test the robustness of our method, we applied it to reading data from two languages, English and Hindi, that differ typologically in several ways, reflecting their distinct linguistic origins, families, and structures. Our work has value in attempting to model probabilistic integration explicitly, as an additional important cognitive mechanism underlying language processing that is currently underappreciated in psycholinguistic modeling.

## 2 Model Architectures

We evaluate two Recurrent Neural Network architectures: a vanilla Recurrent Neural Network (RNN, Elman, 1990), and a variant that introduces a processing delay (d-RNN, Turek et al., 2020). The former has a long tradition in cognitive modeling (Elman, 1990; Rohde and Plaut, 1997; Christiansen and Chater, 1999; Cartling, 2008), as it is naturally suited to implement forward prediction over sequential inputs. The latter was first proposed in the context of NLP, to incorporate sensitivity to backward dependencies (i.e., approximating bidirectional RNNs); however, it has also been used to model language acquisition (Alhama et al., 2021).

RNNs are forward models by design because they are trained to predict the upcoming word in a sentence ($x_{t+1}$) based on two sources of information: the current word $x_t$ and the hidden state of the network, computed in the previous step ($h_{t-1}$). The d-RNN implements next-word prediction in the same way, but its key feature is the addition of a processing delay $d$ such that, for an input word $x_t$, its output is produced at time $t + d$ (i.e. the predicted word is $\hat{y}_{t+d}$). Thus, the weights of the d-RNN are only updated after $d$ extra words, delaying learning. Turek et al. (2020) showed that a large enough delay approximates bidirectional processing, suggesting that the delay allows the network to capture backward dependencies. Importantly, while bidirectional models do exploit context to the left and right of a target word to be predicted, they appear unsuited as cognitive models of real-time incremental language processing, as they require entire sentences or paragraphs to compute their predictions. Instead, and crucially, the d-RNN combines classic forward prediction for incremental input with sensitivity to backward dependencies, making it a suitable cognitive model of prediction *and* integration.

## 3 Data

**Sources.** The choice of English and Hindi is based on three criteria. First, we required languages with different word order, to ensure enough variability in forward and backward dependencies. While English is strictly SVO, Hindi favours SOV order. Second, we chose languages that differ in terms of morphological typology. Our statistical analysis is done at the word level, so we use this criterion to ensure the languages are comprised of words that cannot easily be separated into multiple morphemes, as in agglutinative languages. English is an analytic language that uses specific words rather than inflection to express syntactic relations. Generally, this entails having one morpheme per word. Hindi is a fusional language: it 'fuses' morphemes together in a word where it is not easy to distinguish the individual morphemes (Ramoo, 2021).

Thirdly, for model training and validation against human reading patterns (specifically, eye fixations on words), we sourced publicly accessible datasets for each language, containing texts of a uniform style. We utilized the Potsdam-Allahabad Hindi Eye-tracking Corpus (PAC, Husain et al., 2015; Vasishth, 2021) comprising word-level eye-tracking data from 30 individuals reading 83 sentences sourced from newspapers. For English, our source was the Multilingual Eye-tracking Corpus (MECO, Siegelman et al., 2022), which includes eye-tracking data captured from 46 participants reading 112 encyclopedic sentences. Both human reading datasets align with expository writing, prompting us to select Wikipedia articles for training our language models. These articles provide a congruent encyclopedic text style and are widely available across languages from Wikimedia Foundation dumps.

**Pre-processing.** We pre-process the Wikipedia articles and the PAC and MECO sentences using the Stanza library (Qi et al., 2020) for tokenization, Part-of-Speech annotation and lemmatization. We use lowercased text and remove all punctuation (but we process sentences separately). In the case of Wikipedia texts, we remove article titles using regular expressions, and randomly sample 200,000 sentences for each language. The chosen corpora appear comparable in their mean sentence

lengths: 18.73 and 21.07 words-per-sentence (wps) for the English training set and MECO sentences, respectively, and 18.94 and 16.2 wps for the Hindi training set and PAC sentences, respectively.

We introduce an unknown token to handle out-of-vocabulary (OOV) and rare words. When setting low frequency words in the corpus to the unknown token, we are reducing the vocabulary size the RNNs need to train on. This simplifies the task, reducing training time (Chen et al., 2019), but it also reduces the amount of text in MECO and PAC sentences. We find that a cut-off word frequency of 24 leaves us with almost 90% of the data within MECO and PAC, while the vocabulary of the training set in English and Hindi is reduced to some 11 thousand words.

The type-token ratio ($TTR$) calculated after the vocabulary size reduction shows MECO and PAC have a higher degree of lexical variety than the RNN training sets[1]. This is as expected since these datasets contain fewer sentences compared to the amount sampled from Wikipedia. Moreover, the similarity of the training sets' $TTR$s indicates that the RNNs' task difficulty is similar across languages.

Finally, MECO and PAC require minor preprocessing. In both human reading datasets, we remove skipped words, as they do not help us quantify the predictability of a word and the processing effort required to read it. The code used for preprocessing and the rest of our research is available at `https://github.com/ninadelcaro/predict-integrate-cmcl`.

## 4 Experimental Setup

**Language Models.** We use the code for the RNN and d-RNN by Alhama et al. (2021), with a delay of 1 for the d-RNN. Both networks have three layers: an embedding layer, a recurrent one, and a fully connected layer with softmax activation. We feed the networks with the tokenized sentences described above, and we use cross-entropy loss on next-word prediction objective. We update the weights with Stochastic Gradient Descent. We train until the loss becomes stable (around 45 epochs) in the classic RNN and use the same number of epochs to train the d-RNN (41 epochs for English and 45 for Hindi).

Hyperparameter optimization is done using random search (Bergstra and Bengio, 2012). We train on 80% of the Wikipedia articles, using 10% as the validation set and the other 10% as the testing set. The hyperparameters we optimize are the word embedding and hidden state dimensions as well as the learning rate. We select the RNN model with the lowest loss on the validation set and make sure there is no overfitting by comparing the validation loss to the training loss. Our final model has a hidden state size of 682, an embedding size of 426, and a learning rate of 0.001. We use these same hyperparameters across all models (i.e., for both RNN variants and languages).

**Predictor Variables from Language Models.** Following established computational psycholinguistics literature, we use per-word information-theoretic measures of entropy and surprisal (Hale, 2016). Word surprisal is the negative log-probability of said word, and it intuitively quantifies its unexpectedness. This measure has been linked to human sentence processing difficulty and is predictive of eye movements (Levy, 2008; Wilcox et al., 2023b; Aurnhammer and Frank, 2018; Boston et al., 2008; Ehrlich and Rayner, 1981; Merkx and Frank, 2021; Oh and Schuler, 2023; Demberg and Keller, 2008; Smith and Levy, 2013; Shain et al., 2020). Entropy, on the other hand, quantifies the degree of uncertainty over possible outcomes (Shannon, 1948), and it has also been shown to correlate with human sentence processing effort (Keller, 2004; Linzen and Jaeger, 2014; Wilcox et al., 2023b; Hale, 2003; Linzen and Jaeger, 2016; Roark et al., 2009). We compute these metrics for each word in the MECO and PAC datasets, using the probability distributions predicted by our language models.

**Outcome Variables: Eye gazes while reading.** Metrics of reading processing difficulty available from MECO and PAC include: first fixation duration, first-pass reading time, and total fixation time. Because no consensus exists on whether these measures underlie separate cognitive processes, these reading times (RTs) were used as dependent variables in separate regression models (Agrawal et al., 2017; Boston et al., 2008; Keller, 2004; Merkx and Frank, 2021). RTs were log-transformed for normalization, variance stabilization, and outlier influence reduction (Aurnhammer and Frank, 2018).

**Statistical Inference Model.** As in previous research (Agrawal et al., 2017; Aurnhammer and

[1]MECO: $TTR = .34$; PAC: $TTR = .25$; English training set: $TTR = .0023$; Hindi training set: $TTR = .0022$

Frank, 2018; Boston et al., 2008; Merkx and Frank, 2021), we analyze the relationship between the per-word information-theoretic metrics of entropy and suprisal from our language models as predictors, and human reading times as outcomes, using Generalised Linear Mixed Effects Regression models (GLMER) that incorporate both fixed and random effects. The hierarchical structure of MECO and PAC reading data makes the word-level observations non-independent, because the sentences contain words that are embedded in sentences that are read by specific participants. Therefore, we require random intercepts for the participant and the word read to be part of the linear regression models.

We use nested modelling to compare GLMER models with additional independent variables to a baseline GLMER model. Besides random effects, the baseline regresses the eye-tracking data on these control fixed effects covariates known to affect reading times: word length, and order of appearance of each word within the sentence presented to the reader. Model comparison is performed with a log-likelihood ratio test, allowing us to test whether a single predictor added at each step explains any more variance in the outcome variable by improving the model fit.

Entropy and surprisal are not correlated (Pearson's $r(119306) = 0.36, p < 0.001$) in the RNN regression, but they are in the d-RNN regression (Pearson's $r(119306) = 0.99, p < 0.001$). Therefore, we choose to separate these two metrics in two sets of stepwise GLMER models, one using entropy and another using surprisal as predictor variables. Each set consists of a) the baseline model, b) a model adding the RNN's metric, and c) a model adding the d-RNN's metric to the previous model. We can thus rigorously evaluate our key theoretical conjecture: does the d-RNN architecture, which incorporates a form of language integration, contribute incremental variance over and above an RNN that operates solely on a predictive mechanism?

## 5 Results

Table 1 presents the core outcomes of our GLMER analysis, with detailed model comparisons, log-likelihood ratio tests, and $\alpha$ significance levels provided in Appendix A.

**English.** For the English reading dataset, word entropy of the RNN did not improve the baseline model for any of the three dependent variables, whereas the d-RNN's entropy did so when considering first fixation and total fixation duration as dependent variables. Conversely, adding the surprisal of the RNN improved model fit for all three dependent variables, while adding the surprisal of the d-RNN did not improve model fit further.

| | | English | | Hindi | |
|---|---|---|---|---|---|
| **Metric** | **Model** | **Ent.** | **Surp.** | **Ent.** | **Surp.** |
| FFD | RNN | .39 | .002 | .02 | .11 |
| | d-RNN | .01 | .13 | .1 | .19 |
| TFD | RNN | .22 | <.001 | <.001 | <.001 |
| | d-RNN | <.001 | .07 | .02 | .3 |
| FPRT | RNN | .8 | .002 | .02 | .08 |
| | d-RNN | .08 | .56 | .02 | .05 |

Table 1: Nested model comparison results for human reading time outcomes. Each model comprises various predictors—RNN Model with baseline predictors plus RNN metric, and d-RNN Model with added d-RNN metric. The table shows *p-values* from log-likelihood ratio tests for model comparisons. FFD: First Fixation Duration; TFD: Total Fixation Duration; FPRT: First Pass Reading Time; Ent.: Entropy; Surp.: Surprisal.

**Hindi.** In the Hindi reading dataset, adding the RNN's word entropy to the baseline model improved the model, and so did adding the d-RNN's entropy when predicting total fixation duration and first pass reading time. On the other hand, model comparison revealed no model fit improvement when entering word surprisal, with a notable exception: the addition of the RNN's surprisal to the model regressing total fixation duration.

## 6 Discussion

The ephemeral nature of language is evident, as it rapidly vanishes from our sensory experience upon its completion – being spoken or read. While current psycholinguistics research primarily emphasizes probabilistic prediction as a mechanism that facilitates efficient language learning and real-time processing, the computational modeling of integration and its interplay with prediction in human sentence processing remain less understood. Addressing this, we used an RNN to model pure prediction and a d-RNN for the combined processes of prediction and integration, and assessed the relationship between language model-derived entropy and surprisal measures and eye-tracking data.

The d-RNN's entropy contribution across languages suggests that language models incorporat-

ing integration explain variability in eye-tracking data beyond prediction alone, although surprisal did not yield similar results. A tentative interpretation is that the time course of integration is better reflected in a metric like entropy, which measures uncertainty based on the current state of knowledge of the model, rather than in an a-posteriori and word-specific metric like surprisal. This may be a consequence of the specific operationalization of integration provided by the d-RNN, which delays learning until subsequent words have been processed. Such operationalization is in fact reminiscent of the *lookahead* mechanism used in the parsing literature, which peeks at a number of upcoming tokens in a sentence in order to decide between alternative syntactic analyses (Marcus, 1980; Stabler, 1983; Nozohoor-Farshi, 1986).

The different outcomes in English and Hindi data could suggest that integration and prediction may be employed differently in various languages, possibly influenced by the distinct word orders of the languages we examined—English being SVO and Hindi SOV— and how they interact with RNN model metrics and eye-tracking measures. These observations call for additional investigations into a broader spectrum of languages to discern how language structure might tip sentence processing toward either integration or prediction.

Note that in modeling reading processes, we strived for cognitive plausibility. While more recent and powerful architectures such as bidirectional recurrent networks and encoder-decoder transformers could potentially implement integration, they also do it using text from the future, i.e. they require entire sentences or passages to predict a masked word and train its algorithm. Since relying on future words is not cognitively plausible when processing language word-by-word incrementally, we opted for classic RNN implementations. Other models like Long-Short Term Memory Networks (Hochreiter and Schmidhuber, 1997) and decoder-only transformers trained unidirectionally (Radford et al., 2019) meet our requirements, and we leave the investigation of their suitability to future work.

## Acknowledgments

We thank Samar Husain and Shravan Vasishth for additional information on the PAC dataset and appreciate Tilburg University's GPU4EDU project for the GPU resources provided.

## References

Arpit Agrawal, Sumeet Agarwal, and Samar Husain. 2017. Role of expectation and working memory constraints in hindi comprehension: An eye-tracking corpus analysis. *Journal of Eye Movement Research*, 10.

Raquel G. Alhama, Francesca Zermiani, and Atiqah Khaliq. 2021. Retrodiction as delayed recurrence: the case of adjectives in Italian and English. *Proceedings of the 19th Workshop of the Australasian Language Technology Association*, pages 163–168.

Christoph Aurnhammer and Stefan Frank. 2018. Comparing gated and simple recurrent neural network architectures as models of human sentence processing. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

James Bergstra and Y. Bengio. 2012. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13:281–305.

Marisa Ferrara Boston, John Tracy Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, 2.

Bo Cartling. 2008. On the implicit acquisition of a context-free grammar by a simple recurrent neural network. *Neurocomputing*, 71(7):1527–1537.

Wenhu Chen, Yu Su, Yilin Shen, Zhiyu Chen, Xifeng Yan, and William Yang Wang. 2019. How large a vocabulary does text classification need? a variational approach to vocabulary selection. 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)*, pages 3487–3497, Minneapolis, Minnesota. Association for Computational Linguistics.

Morten H Christiansen and Nick Chater. 1999. Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23(2):157–205.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Susan F. Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Fernanda Ferreira and Suphasiree Chantavarin. 2018. Integration and prediction in language processing: A synthesis of old and new. *Current Directions in Psychological Science*, 27:096372141879449.

Wikimedia Foundation. Wikimedia Downloads.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

John Tracy Hale. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32:101–123.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Samar Husain, Shravan Vasishth, and Narayanan Srinivasan. 2015. Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research*, 8:1–12.

Frank Keller. 2004. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 317–324, Barcelona, Spain. Association for Computational Linguistics.

Gina R Kuperberg and T Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, 31(1):32–59.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Tal Linzen and Florian Jaeger. 2014. Investigating the role of entropy in sentence processing. In *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*, pages 10–18, Baltimore, Maryland, USA. Association for Computational Linguistics.

Tal Linzen and T. Florian Jaeger. 2016. Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, 40(6):1382–1411.

Mitchell P. Marcus. 1980. *Theory of Syntactic Recognition for Natural Languages*. MIT Press, Cambridge, MA, USA.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22, Online. Association for Computational Linguistics.

Mante S Nieuwland, Dale J Barr, Federica Bartolozzi, Simon Busch-Moreno, Emily Darley, David I Donaldson, Heather J Ferguson, Xiao Fu, Evelien Heyselaar, Falk Huettig, et al. 2020. Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B*, 375(1791):20180522.

R. Nozohoor-Farshi. 1986. On formalizations of Marcus' parser. In *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*, pages 533–535.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Luca Onnis and Falk Huettig. 2021. Can prediction and retrodiction explain whether frequent multi-word phrases are accessed 'precompiled' from memory or compositionally constructed on the fly? *Brain Research*, 1772:147674.

Luca Onnis, Alfred Lim, Shirley Cheung, and Falk Huettig. 2022. Is the mind inherently predicting? exploring forward and backward looking in language processing. *Cognitive Science*, 46(10):e13201.

Tiago Pimentel, Clara Meister, Ethan Wilcox, Kyle Mahowald, and Ryan Cotterell. 2023. Revisiting the optimality of word lengths. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2240–2255, Singapore. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Dinesh Ramoo. 2021. Morphology of different languages. In *Psychology of Language*. BCcampus.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333, Singapore. Association for Computational Linguistics.

Douglas L. T. Rohde and David C. Plaut. 1997. Simple recurrent networks and natural language: How important is starting small?

Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. 2020. fmri reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307.

C. E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Noam Siegelman, Sascha Schroeder, Cengiz Acarturk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Fonseca, Nicolas Dirix,

Wouter Duyck, Argyro Fella, Ram Frost, Carolina Gattei, Areti Kalaitzi, Nayoung Kwon, Kaidi Lõo, and Victor Kuperman. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior Research Methods*, 54:1–21.

Nathaniel J. Smith and R. Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

Edward P. Stabler. 1983. Deterministic and bottom-up parsing in Prolog. In *Proceedings of the Third AAAI Conference on Artificial Intelligence*, AAAI'83, page 383–386. AAAI Press.

Javier Turek, Shailee Jain, Vy Vo, Mihai Capotă, Alexander Huth, and Theodore Willke. 2020. Approximating stacked and bidirectional recurrent architectures with the delayed recurrent neural network. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9648–9658. PMLR.

Shravan Vasishth. 2021. vasishth/lingpsych: Data and Functions used in the Book "Linear Mixed Models in Linguistics and Psychology: A Comprehensive Introduction". *rdrr.io*.

Ethan Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023a. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511, Singapore. Association for Computational Linguistics.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023b. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

# A Appendix

| | | English | | | Hindi | | |
|---|---|---|---|---|---|---|---|
| **Outcome** | **Predictors** | **AIC** | $\chi^2$ | **p-value** | **AIC** | $\chi^2$ | **p-value** |
| First fixation duration | Baseline | 55443 | | | 23932 | | |
| | RNN entropy | 55444 | 0.74 | .39 | 23928 | 5.8 | .02 |
| | d-RNN entropy | 55440 | 6.36 | .01 | 23927 | 2.75 | .1 |
| Total fixation duration | Baseline | 94736 | | | 37375 | | |
| | RNN entropy | 94737 | 1.52 | .22 | 37366 | 11.16 | <.001 |
| | d-RNN entropy | 94727 | 12.08 | <.001 | 37363 | 5.23 | .02 |
| First pass reading time | Baseline | 72201 | | | 32505 | | |
| | RNN entropy | 72203 | 0.06 | .8 | 32502 | 5.84 | .02 |
| | d-RNN entropy | 72202 | 2.97 | .08 | 32499 | 5.06 | .02 |

Table 2: Results of stepwise nested model comparisons predicting human reading time outcomes. Each inference model includes different predictors: Baseline Model (word length, sentence position, and subject and word random intercepts), RNN Entropy Model (Baseline predictors plus RNN word entropy), and d-RNN Entropy Model (RNN Model predictors plus d-RNN word entropy). Models are assessed using log-likelihood ratio tests.

| | | English | | | Hindi | | |
|---|---|---|---|---|---|---|---|
| **Outcome** | **Predictors** | **AIC** | $\chi^2$ | **p-value** | **AIC** | $\chi^2$ | **p-value** |
| First fixation duration | Baseline | 55443 | | | 23932 | | |
| | RNN surprisal | 55436 | 9.22 | .002 | 23931 | 2.54 | .11 |
| | d-RNN surprisal | 55436 | 2.27 | .13 | 23932 | 1.74 | .19 |
| Total fixation duration | Baseline | 94736 | | | 37375 | | |
| | RNN surprisal | 94709 | 28.96 | <.001 | 37360 | 17.7 | <.001 |
| | d-RNN surprisal | 94708 | 3.23 | .07 | 37360 | 1.08 | .3 |
| First pass reading time | Baseline | 72201 | | | 32505 | | |
| | RNN surprisal | 722194 | 9.84 | .002 | 32504 | 3.1 | .08 |
| | d-RNN surprisal | 722195 | 0.34 | .56 | 32503 | 3.72 | .05 |

Table 3: Results of stepwise nested model comparisons predicting human reading time outcomes. Each inference model includes different predictors: Baseline Model (word length, sentence position, and subject and word random intercepts), RNN surprisal Model (Baseline predictors plus RNN word surprisal), and d-RNN surprisal Model (RNN Model predictors plus d-RNN word surprisal). Models are assessed using log-likelihood ratio tests.