Supplementary Materials

[A] Extraction of the learning-based paradigmatic and syntagmatic covariates

We derived paradigmatic and syntagmatic covariates using the Naïve Discrimination Learning (NDL) algorithm, which builds on a theory of learning that is anchored in the equations for error-correction learning proposed by Rescorla and Wagner (1972); these equations are effectively identical to the Widrow-Hoff or Delta rule (Widrow & Hoff, 1960; Rescorla, 2008 explains the relationship). The Rescorla-Wagner rule is also related to the Perceptron (Rosenblatt, 1962), but it is simpler in that it uses total or net input as the strength of activation of an output unit rather than sigmoid squashing or a similar procedure for normalisation. Several studies on language acquisition have shown that the Rescorla-Wagner model is predicts human behaviour well across various language-related tasks and problems (e.g., Ellis, 2006a; Ellis, 2006b; Baayen et al., 2011; Arnon & Ramscar, 2012; Ramscar et al., 2013; Milin et al., 2017; Divjak et al., 2020).

The Rescorla-Wagner rule learns to associate learning cues and outcomes iteratively, in discrete time steps when new information is presented. If we let C denote the set of k possible input cues, and let O denote the set of n possible outputs, a Rescorla-Wagner learning network will be defined by an $k \times n$ associative weight matrix, given k cues and n outcomes. Weights are adjusted or updated for each discrete learning event on which a subset of all possible cues and outcomes will be present. The weights update is defined as:

$$w_i^{t+1} = w_i^t + \Delta w_i^t$$

with the change in weight, Δw_i, which updates the weight, w_i as new information arrives ($t + 1$). How the change in weight, Δw_i is calculated, is defined by the Rescorla-Wagner equations:

(1)	The cue is ABSENT	**nothing** happens	$\Delta w_i^{t+1} = 0$
(2)	The cue is PRESENT; The outcome is PRESENT	**positive evidence** that should strengthen the connection weight	$\Delta w_i^{t+1} = \gamma\left(1 - \sum w_i^t c_i\right)$
(3)	The cue is PRESENT; The outcome is ABSENT	**negative evidence** that should weaken the connection weight	$\Delta w_i^{t+1} = \gamma\left(0 - \sum w_i^t c_i\right)$

The weight update depends on all cues, c_i, present at the event in time t, and the *learning rate*, γ, which is the only free-parameter (compare Rescorla & Wagner, 1972 vs. Widrow & Hoff, 1960 or see the summary in Milin et al., 2020). The error that is corrected iteratively is defined as the difference between the target or the truth – the outcome is present (1) or absent (0) – and the current state of knowledge, which is the weighted sum of present cues ($\sum w_i^t c_i$).

For the present training, we made use of the Rescorla-Wagner rule implementation in the **pyndl** library (Sering et al., 2017) for **Python** (v. 3.x), and data from the srWaC corpus (Ljubešić & Klubička, 2016). The data pre-processing included removing punctuation, all non-Serbian characters (to filter out non-Serbian words), and lowercasing all words. The final dataset for training included 780,404,835 word tokens across 40,878,185 sentences.

Two models were trained. The orthographic model (G2F) was trained on letter triplets and used chunks of three consecutive words as learning event; all attested letter triplets served as input cues and the word in the middle served as outcome. The lexical model (F2F) was trained on words and used two preceding words and one following word as learning events, with the target word as the third word in the sequence (a moving window). For example, for the sentence “One of these days all the boys and girls will finally meet” a typical G2F learning event would be “and girls will”, with letter triplets (trigraphs) as orthographic cues “#an, and, nd#, d#g, #gi, gir, irl, rls, ls#, s#w, #wi, wil, ill, ll#” and with the middle word “girls” as the outcome (note that the hash symbol replaces word boundaries). For the F2F lexical model, an event would be “boys and girls will”, with the outcome again being “girls”, and with cues “boys and” (the two preceding words) and "will" (the one following word). In cases where an outcome would occupy the initial position in a sentence, no preceding context would be available; similarly, the final word in a sentence would lack learning cues that would follow that word. To follow up on the example above, small sections of the two resulting matrices, with cues in rows and outcomes in columns, could look as follows:

	boys	girls	finally	time	
#an	-0.016	-0.016	-0.062	-0.009	×××
and	0.051	0.022	0.037	-0.027	×××
nd#	-0.059	-0.002	0.006	0.039	×××
	×××	×××	×××	×××	×××
ill	-0.018	0.065	0.019	-0.061	×××
ll#	-0.001	-0.034	-0.016	-0.058	×××
	×××	×××	×××	×××	×××

	boys	girls	finally	time	
boys	0.000	0.003	-0.002	0.004	×××
girls	-0.001	0.000	-0.009	0.000	×××
finally	-0.003	0.009	0.000	0.008	×××
	×××	×××	×××	×××	×××
time	0.000	0.002	-0.008	0.000	×××
age	-0.002	0.005	-0.008	0.007	×××
	×××	×××	×××	×××	×××

On the left, we see a portion of the orthographic learning model (G2F), trained on letter triplets as input cues and word forms as outcomes. This model provides two indicators of paradigmatic relationships: TrigraphActivation, a measure of orthographic support for a target word, and TrigraphCompetition, a measure of competition among orthographically similar words. For an example sequence of three words (“and girls will”) with their respective letter triplets (“#an, and, nd#, d#g, #gi, gir, irl, rls, ls#, s#w, #wi, wil, ill, ll#”), TrigraphActivation is calculated as the sum of weights in the target word's column (“girls”) across the 14 input cues present. In contrast, TrigraphCompetition sums the absolute values of all weights in the rows corresponding to the same 14 cues. In essence, TrigraphCompetition quantifies orthographic competition as the absolute vector length (1-norm), capturing the extent to which other words are relevant given the visual input.

The lexical learning model (F2F) is illustrated on the right. Here, the matrix yields six potential measures of syntagmatic relationships. ActivationToContext reflects the activation the target word provides to other words in context, analogous to *word2vec*’s Skip-Gram, and is calculated as the 1-norm of the target word’s row (a cue). ActivationFromContext, similar to *word2vec*’s CBOW, is the 1-norm of the target word’s column (an outcome). The remaining four predictors capture *relational* properties. Two of these, SimilarityToContext and SimilarityFromContext, indicate the relationships between the prime and target nouns. They are quantified as cosine similarity between row- and column-based vectors for the respective prime and target nouns. The final two predictors, TypicalityToContext and TypicalityFromContext, measure how (non)exceptional the target word is compared to an "average word". These are calculated as cosine similarity between the target noun's row and column vectors and the weight-average vectors across all rows and columns, respectively.

NOTE: Large pretrained NDL matrices for English and Polish are freely available at https://outofourminds.bham.ac.uk/cloudcomputing/. Extracting static word embeddings (vectors of 50–1000 elements) is straightforward, as explained on the website, and additional languages will be added.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*(3), 292-305.

Baayen, R. H., Milin, P., Filipović Đurđević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review*, *118*(3), 438-481. https://doi.org/10.1037/a0023851

Divjak, D., Milin, P., Ez-Zizi, A., Józefowski, J., & Adam, C. (2020). What is learned from exposure: an error-driven approach to productivity in language. *Language Cognition and Neuroscience*, *36*(1), 60-83.

Ellis, N. C. (2006a). Language acquisition as rational contingency learning. *Applied linguistics*, *27*(1), 1-24.

Ellis, N. C. (2006b). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied linguistics*, *27*(2), 164-194.

Ljubešić, N., & Klubička, F. (2016). *The Serbian web corpus srWaC*

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PloS one*, *12*(2), e0171935.

Milin, P., Madabushi, H. T., Croucher, M., & Divjak, D. (2020). Keeping it simple: Implementation and performance of the proto-principle of adaptation and learning in the language sciences. *arXiv preprint arXiv:2003.03813*.

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of" mouses" in adult speech. *Language*, 760-793.

Rescorla, R. A. (2008). Rescorla-Wagner model. *Scholarpedia*, *3*(3), 2237.

Rescorla, R. A., & Wagner, R. A. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In H. Black & W. F. Proksay (Eds.), *Classical conditioning II* (pp. 64-99). Appleton-Century-Crofts.

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and*
the Theory of Brain Mechanisms. Spartan Books.

Sering, K., Weitz, M., Künstle, D.-E., & Schneider, L. (2017). *pyndl: Naive discriminative learning in Python*. In https://doi.org/10.5281/zenodo.597964

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. WESCON Convention Record Part IV,