Retrodiction as Delayed Recurrence: the Case of Adjectives in Italian and English

Raquel G. Alhama[1,2] **Francesca Zermiani** [2,3] **Atiqah Khaliq**[2]

[1]Tilburg University, Department of Cognitive Science & Artificial Intelligence, Tilburg, The Netherlands
[2]Max Planck Institute for Psycholinguistics, Language Development Department, Nijmegen, The Netherlands
[3]University of Stuttgart, Department of Teaching and Learning with Intelligent Systems, Stuttgart, Germany
rgalhama@tilburguniversity.edu,
francesca.zermiani@ife.uni-stuttgart.de, atiqah.khaliq@mpi.nl

Abstract

We address the question of how to account for statistical dependencies in an online processing account of human language acquisition. We focus on descriptive adjectives in English and Italian, and show that the acquisition of adjectives in these languages likely relies on tracking both forward and backward regularities. Our simulations confirm that forward-predicting models like standard Recurrent Neural Networks cannot account for this phenomenon due to the lack of backward prediction, but the addition of a small delay (as proposed in Turek et al., 2019) endows the RNN with the ability to not only predict but also retrodict.

1 Introduction

Sensitivity to statistical regularities allows for efficient lexical processing. As a sentence unfolds, the experienced words convey information that humans use to anticipate upcoming words, and gain thereby processing speed. This has been evidenced in a long tradition of studies with human reading data, which reveal that words that are more predictable given their context are more likely to be read faster or even skipped (Ehrlich and Rayner, 1981).

The ability to track statistical regularities during language processing is present at a very young age, and can be recruited for language learning. Before their first birthday, infants are able to use this skill to identify words in unknown languages created with artificial (Saffran et al., 1996; Aslin et al., 1998) or natural words (Pelucchi et al., 2009b), demonstrating that this ability is useful for learning language-like stimuli (see Saffran, 2020 for a review). Studies have found that, before their second year of age, toddlers already engage in predictive processing to identify familiar words before they are complete (Swingley et al., 1999; Fernald et al., 2001), and are capable of anticipating upcoming words (Fernald and Hurtado, 2006; Lew-Williams and Fernald, 2007).

Given this relation between online processing and learning, it is perhaps unsurprising that children with more efficient lexical processing are also those with faster vocabulary growth (Fernald et al., 2006; Fernald and Marchman, 2012; Weisleder and Fernald, 2013; Donnelly and Kidd, 2020). From a cross-linguistic perspective, this suggests that typological variation on the statistical regularities of different languages should be either equally tracked during processing, or reflected in cross-linguistic differences in learning.

In our work, we focus on one such typological feature: in particular, word order of descriptive adjectives in English (which occur pre-nominally), and Italian (which appear mostly post-nominally, but also pre-nominally). We first show that this difference in word order bears a different pattern of statistical dependencies in these languages, related to the direction in which the words in these constructions are more predictable (forward in Italian, backward in English). We find that, despite this difference, children acquire nouns and adjectives in each language at the same pace, showing no advantage of either direction. Thus, in line with the relation between processing and learning sketched above, a computational approach needs to accommodate statistical tracking of dependencies that are both forward and backward. We then show limitations of standard recurrent models in dealing with backward dependencies, and propose the use of a Delayed Recurrent Neural Network (Turek et al., 2019) to capture this phenomenon.

2 Related work

Word order constraints in languages may favor regularities in either the *forward* direction (when words are predictable from their *earlier* context) or in the *backward* direction, (when words are more predictable from the context occurring *after* them). For instance, languages differ in whether they use prepositions (e.g. 'in Paris') or postpositions (e.g. 'Paris in'). The conditional probability of observing 'in' given 'Paris' is higher than that of observing 'Paris' given 'in' (since 'in' may be preceded or followed by any location name); thus, the construction with a postposition has higher forward predictability, while the construction with a preposition is more predictable in the backward direction.

Few studies focus on the role of backward dependencies in human processing and learning. Pelucchi et al. (2009a) found that 8-month-old infants can learn the words of an artificial language which can only be identified based on conditional probabilities in the backward direction. The experiments reported in Perruchet and Desaulty (2008) demonstrate that this ability is also present in adults. Another set of studies revealed that the word order patterns in the native language of speakers create learning biases that manifest when learning an artificial language. Using a carefully controlled artificial language that contained balanced cues in the forward and backward direction, Onnis and Thiessen (2013) found a significant difference between Korean and English speakers, manifested in a tendency to rely on dependencies that are consistent with the direction that best predicts constituency in those languages (forward for Korean and backward for English). 13 month-old children learning English also exhibit this bias (Thiessen et al., 2019).

French et al. (2011) reported successful simulations of the experiments in Perruchet and Desaulty (2008) with an autoencoder. This model used a form of recurrence that was conditioned on the reconstruction error, such that only internal representations of items with low error would be fed back to the model on the next step. Simulations involving standard recurrence were not successful in learning the backward dependencies (Perruchet and Peereman, 2004).

3 Corpus Analysis

First we confirmed that the adjective order in English and Italian was reflected in the conditional probabilities between adjectives and nouns. We extracted child-directed speech transcriptions from all the English and Italian corpora available in CHILDES (MacWhinney, 2000), using the `childesr` library (Sanchez et al., 2019) [1]. We focused on ages from 0 to 60 months old, and used the lemmatized, lowercased version of the words. Since part-of-speech information was not available for all the data, we used the part-of-speech tagger in spaCy [2] to annotate it. We applied additional manual revision to remove some words that were wrongly classified as descriptive adjectives [3]. We used the lemmatized version of the words since, unlike in English, nouns and adjectives have number and grammatical gender in Italian.

We selected all the adjective-noun pairs (for both languages), and noun-adjective pairs (for Italian only). We downsampled the adjective-noun pairs in English to be comparable in size to the Italian data. For each word pair $w_1 w_2$ we computed its conditional probability as $P(w_i|w_j) = counts(w_1 w_2)/counts(ctx)$, where $i = 2, j = 1, ctx = w_1$ for forward conditional probabilities and $i = 1, j = 2, ctx = w_2$ for backward conditional probabilities.

Figure 1 shows the distribution of the computed probabilities. Whereas forward conditional probabilities are significantly more reliable for adjectives occurring in the Italian canonical noun-adjective ordering ($p < 0.01$), the opposite is the case for English, in which predicting backwards is significantly more reliable ($p < 0.001$). In the case of the adjective-noun order in Italian, both forward and backward probabilities are equally informative. This is consistent with the highly formulaic nature of this syntactic pattern, since not all adjectives and nouns occur in this construction. To summarize, as expected, word order is reflected in the conditional probabilities between adjectives and nouns, at least in the canonical order: while noun-adjective in Italian is favoured by forward probabilities, adjective-noun in English is better predicted backwards.

[1] `http://childes-db.stanford.edu`

[2] `https://spacy.io`. Models: `it_core_news_sm` and `en_core_web_sm`.

[3] All the code used for data processing, analyses and models is available at `https://github.com/rgalhama/retro_adjs`.

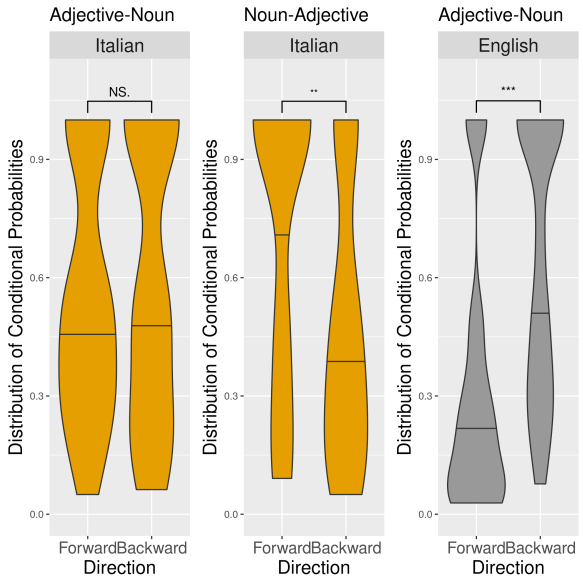

Figure 1: Distribution of conditional probabilities between words in adjective-noun and noun-adjective pairs, for English and Italian. Asterisks indicate if p-values are under significance levels (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; N.S.: $p > 0.05$).

4 The Acquisition of Adjectives

Efficient processing correlates with faster vocabulary growth. Thus, if there is a difference in processing forward and backward dependencies, it should be reflected in a cross-linguistic difference in vocabulary acquisition (note that, once children start producing adjectives, they rarely produce them in incorrect word order, (Nicoladis, 2006)).

To analyze this, we used data collected with the MacArthur-Bates Communicative Development Inventory forms (CDIs). These forms contain checklists of common early acquired words. Parents complete the forms according to their estimation of whether their child produced each of those words at a given age. We used the 'Words & Sentences' CDIs from Wordbank (Frank et al., 2017)[4], for English and Italian. We excluded the forms involving twins (as significant differences have been observed in the language development of twins and singletons, Tomasello et al., 1986). We used the library in Wordbank to estimate the age of acquisition (AoA), considering that a word is acquired at the age at which at least 50% of the children in the sample produced a given word. Since differences in the acquisition of nouns could have an effect on the AoA of adjectives, we also report the estimated AoA of nouns.

[4] http://wordbank.stanford.edu/

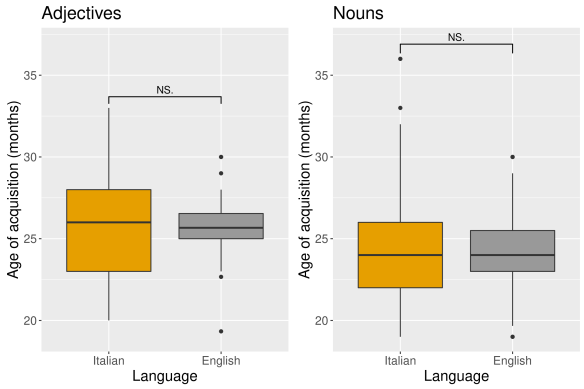

Figure 2: Age of Acquisition (AoA) of adjectives and nouns, as estimated from the CDIs in Wordbank.

As can be seen in Fig. 2, there is no significant difference between the AoA of adjectives and nouns in each language, even though we find more variability in Italian. This result suggests that children learning Italian must be employing their forward predictive skills, while children learning English need to draw upon their capacity to retrodict.

5 Do RNNs Retrodict?

To account for the results in the previous section, models of online processing should predict but also retrodict. We first present simulations with a Recurrent Neural Network (RNN, Elman, 1990), which has a long tradition of use as a model of human sequential processing (with equivalent performance to variants with gated recurrence Aurnhammer and Frank, 2019). Although the RNN is trained exclusively in the forward direction, it is necessary to rule out the possibility that it can implicitly learn patterns that capture the backward regularities.

We trained the RNN on the child-directed data described in section 3, including sentences with and without adjectives. We downsampled the English data to have comparable training data size (41862 sentences). The RNN had an embedding layer (size:100), a hidden recurrent layer (size:250), and a softmax output layer over the whole vocabulary (size: 7875 (English); 7520 (Italian)). The model was trained to predict the next word in a sentence. We used cross-entropy loss, and updated the weights of the model with Stochastic Gradient Descent, until the loss became stable (around 60 epochs). We evaluated the trained model based on the entropy of the model prediction after the first word in adjective construction. Results are shown in Figure 3.

As can be seen, at the end of training, the RNN

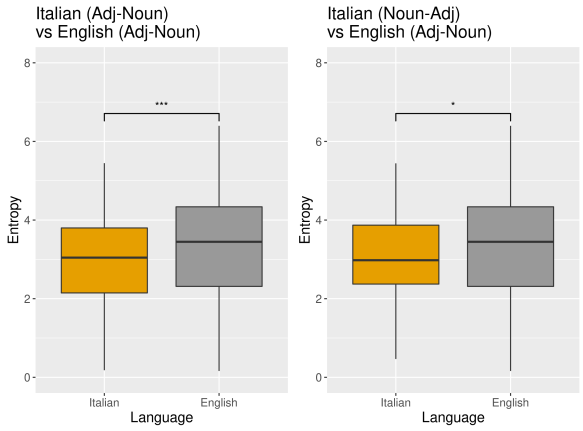

Figure 3: Entropy at the output layer of the RNN, after the first word in each adjective-noun or noun-adjective pair. Asterisks indicate if p-values are under significance levels (*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; N.S: $p > 0.05$).

is significantly less successful in learning English than Italian (where success is quantified by low entropy). These results are consistent with our expectation: the model performs significantly worse for English, which —as shown in our analyses of conditional probabilities (section 3)— is less favoured by forward probabilities in the adjective-noun construction.

6 Retrodiction as Delayed Prediction

Our results indicate that a strictly forward model like the standard RNN cannot account for learning backward dependencies. An enhancement that could potentially capture the backward dependencies is the addition of a bidirectional recurrent layer (biRNN, Schuster and Paliwal, 1997). However, this would not constitute a realistic account of human processing, as this model peeks into the context which is not yet experienced.

Thus we explore an alternative account of retrodiction that functions as delayed prediction, based on the model presented in Turek et al. (2019), known as Delayed Recurrent Neural Network (dRNN). The dRNN extends the standard RNN in the following way. In the RNN, when an input word w_t is presented at time t, the model predicts the next word w_{t+1}, and the weights are updated immediately. In the dRNN, the weight update is performed at time $t + d$, where d is the pre-defined 'delay'. This entails that d extra words have been processed by the network before the error is backpropagated. This prevents the model from seeing future words during prediction, but it can effectively see them before the parameter update.

We implement a dRNN with the same hyperparameters as the RNN. We set a delay of one word and evaluate the model with the same entropy measure after similar number of epochs as the RNN (60 epochs). Results are shown in Fig. 4. As can be seen, there are no significant differences between these languages, suggesting that this model can account for learning adjective constructions in both languages.

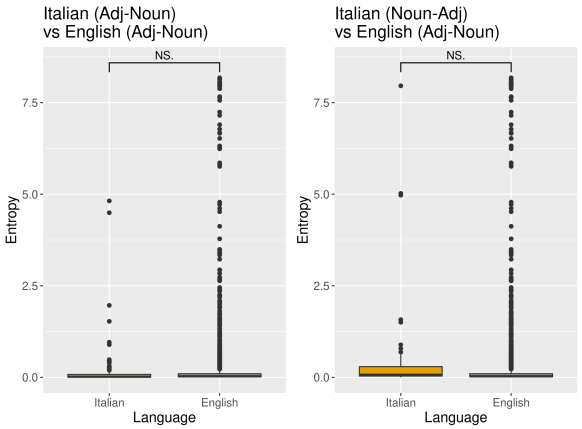

Figure 4: Entropy of the dRNN with $d = 1$, after the first word in each adjective-noun or noun-adjective pair. NS. (Not Significant) indicates p-value> 0.05.

Turek et al. (2019) noted that, for a large enough d, the dRNN can approximate the behavior of a biRNN. Since the biRNN explicitly processes the context after a word in the backward direction, similar performance provides further indication that the dRNN is learning backward dependencies. We thus replicate our simulations with a biRNN. Table 1 summarizes the mean entropy for all the models. As can be seen, the biRNN and the dRNN perform almost identically.

	RNN	biRNN	dRNN
ita: n-adj	3.01(1.08)	0.45(0.94)	0.42(1.20)
ita: adj-n	2.94(1.14)	0.15(0.27)	0.16(0.60)
ita: comb.	2.96(1.12)	0.26(0.61)	0.25(0.87)
eng: adj-n	3.33(1.29)	0.21(0.57)	0.24(0.92)

Table 1: Mean entropy (standard deviation) after the first word in adjectival constructions in Italian (noun-adjective, adjective-noun and both combined) and English (noun-adjective).

This is in line with the reported data, and offers an explanation to why the AoA of children does not show any differences despite the different word order patterns: while a classic RNN account shows

an asymmetry depending on the directions of predictability, by delaying the prediction error update, the dRNN can take advantage of the backward dependencies in English, and strikes a good balance between the two directions in Italian.

7 Conclusions

Our work suggests that a full account of human processing and learning needs to address typological influences on distributional information, which require tracking of both forward and backward statistical dependencies. While we cannot account for these with standard RNN models, the dRNN can capture both forward and backward dependencies, offering a possible explanation for how humans are able to predict but also retrodict.

Acknowledgments

We are grateful to Evan Kidd for discussions and feedback on earlier versions of this paper. The authors also thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Francesca Zermiani.

References

Richard N Aslin, Jenny R Saffran, and Elissa L Newport. 1998. Computation of conditional probability statistics by 8-month-old infants. *Psychological science*, 9(4):321–324.

Christoph Aurnhammer and Stefan L Frank. 2019. Comparing gated and simple recurrent neural network architectures as models of human sentence processing. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

Seamus Donnelly and Evan Kidd. 2020. Individual differences in lexical processing efficiency and vocabulary in toddlers: A longitudinal investigation. *Journal of Experimental Child Psychology*, 192:104781.

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.

Anne Fernald and Nereyda Hurtado. 2006. Names in frames: Infants interpret words in sentence frames faster than words in isolation. *Developmental science*, 9(3):F33–F40.

Anne Fernald and Virginia A Marchman. 2012. Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child development*, 83(1):203–222.

Anne Fernald, Amy Perfors, and Virginia A Marchman. 2006. Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental psychology*, 42(1):98.

Anne Fernald, Daniel Swingley, and John P Pinto. 2001. When half a word is enough: Infants can recognize spoken words using partial phonetic information. *Child development*, 72(4):1003–1015.

Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. 2017. Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, 44(3):677–694.

Robert M French, Caspar Addyman, and Denis Mareschal. 2011. Tracx: a recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological review*, 118(4):614.

Casey Lew-Williams and Anne Fernald. 2007. Young children learning spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, 18(3):193–198.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3 edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Elena Nicoladis. 2006. Cross-linguistic transfer in adjective–noun strings by preschool bilingual children. *Bilingualism: Language and Cognition*, 9(1):15–32.

Luca Onnis and Erik Thiessen. 2013. Language experience changes subsequent learning. *Cognition*, 126(2):268–284.

Bruna Pelucchi, Jessica F Hay, and Jenny R Saffran. 2009a. Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, 113(2):244–247.

Bruna Pelucchi, Jessica F Hay, and Jenny R Saffran. 2009b. Statistical learning in a natural language by 8-month-old infants. *Child development*, 80(3):674–685.

Pierre Perruchet and Stéphane Desaulty. 2008. A role for backward transitional probabilities in word segmentation? *Memory & cognition*, 36(7):1299–1305.

Pierre Perruchet and Ronald Peereman. 2004. The exploitation of distributional information in syllable processing. *Journal of Neurolinguistics*, 17(2-3):97–119.

Jenny R. Saffran. 2020. Statistical language learning in infancy. *Child Development Perspectives*, 14(1):49–54.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Alessandro Sanchez, Stephan C Meylan, Mika Braginsky, Kyle E MacDonald, Daniel Yurovsky, and Michael C Frank. 2019. childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior research methods*, 51(4):1928–1941.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Daniel Swingley, John P Pinto, and Anne Fernald. 1999. Continuous processing in word recognition at 24 months. *Cognition*, 71(2):73–108.

Erik D Thiessen, Luca Onnis, Soo-Jong Hong, and Kyung-Sook Lee. 2019. Early developing syntactic knowledge influences sequential statistical learning in infancy. *Journal of experimental child psychology*, 177:211–221.

Michael Tomasello, Sara Mannle, and Ann C Kruger. 1986. Linguistic environment of 1-to 2-year-old twins. *Developmental Psychology*, 22(2):169.

Javier S. Turek, Shailee Jain, Vy Vo, Mihai Capota, Alexander G. Huth, and Theodore L. Willke. 2019. Approximating stacked and bidirectional recurrent architectures with the delayed recurrent neural network.

Adriana Weisleder and Anne Fernald. 2013. Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological science*, 24(11):2143–2152.