Word order and the learnability of artificial languages

Bob van Tiel (bobvantiel@gmail.com)
Department of Philosophy, Theology, and Religious Studies
Nijmegen, The Netherlands

Fausto Carcassi (fausto.carcassi@gmail.com)
Institute for Logic, Language, and Computation
Amsterdam, The Netherlands

Xiaochen Y. Zheng (xiaochen.zheng@donders.ru.nl)
Donders Institute for Brain, Cognition, and Behaviour
Nijmegen, The Netherlands

Abstract

Languages vary in the way they typically order subject, verb, and object in transitive sentences. Although all six possible word orders are attested, there is great variability in the frequency with which they occur in the languages of the world. Here, we investigate whether this variability is reflected in differences in the learnability of the possible word orders. Thus, we carried out a language learning experiment in which native English speakers had to learn artificial languages with different word orders. The results suggest that there is broad correspondence between the typological frequency of different word orders and their learnability, which supports the hypothesis that there are cognitive and/or communicative factors that are responsible for the bias in the distribution of word orders. We further analyse the data using a novel computational model for simultaneous vocabulary and word order acquisition.

Keywords: syntax; language universal; language learning; artificial language; communicative efficiency

Introduction

Transitive declarative sentences are sentences that contain a subject (S), verb (V), and direct object (O). In English, such sentences typically have SVO word order:

(1) The boy kicked the ball.

However, other languages have different word orders. Indeed, typological research suggests that all six possible orderings of S, V, and O occur in the languages of the world.

At the same time, this typological research also indicates that there is enormous variability in the frequency with which the possible word orders occur. For example, Dryer (2013) reports that, in a sample of 1,376 languages, 189 lacked a dominant word order. Focusing on the 1,187 remaining languages, the vast majority were either SVO (488) or SOV (564). The next most frequent word order was VSO (95). The remaining word orders VOS (25), OVS (11), and OSV (4) were extremely rare (Haider, 2023; Greenberg, 1963).

The skewed distribution of word orders calls into question *why* languages tend to converge on either SVO or SOV word order. Is this merely a historical accident, caused by the fact that the first language allegedly had SOV word order (Givón, 1979; Maurits & Griffiths, 2014)? Or do certain word orders have cognitive and/or communicative benefits that make them more likely to arise and persist in language evolution?

There are at least two observations that suggest that the uneven distribution of word orders is not accidental. First, sign languages generally develop to have SVO or SOV word order, even if they arise within communities that speak languages with other word orders (Sandler, Meir, Padden, & Aronoff, 2005). Second, in experimental settings in which people have to communicate with gestures, they tend to adopt an SVO or SOV ordering, even if their spoken language employs a different word order (Schouwstra & de Swart, 2014; Goldin-Meadow, So, Özyürek, & Mylander, 2008; Futrell et al., 2015).

Both observations indicate that communicators naturally gravitate towards certain word orders over others, suggesting that certain word orders have intrinsic benefits. In the literature, at least three potential benefits have been described.

First, Kemmerer (2012) points out that transitive sentences prototypically denote a causal chain consisting of an agent doing something to a patient. Importantly, this causal chain starts with the agent and ends with the patient. Linguistically, the agent is usually—though not invariably, cf. e.g., passive sentences—referred to by the subject; the patient by the object. Hence, Kemmerer suggests, a possible reason for why, in almost all languages, the subject precedes the object is to mirror the precedence of the agent over the patient in the causal chain (Comrie, 1989; Greenberg, 1963).

Second, Maurits, Perfors, and Navarro (2010) argue that languages almost never start transitive sentences with the object because object-initial transitive sentences are suboptimal in terms of the rate at which information is conveyed. Based on a corpus analysis, Maurits and colleagues show that objects tend to carry more information than subjects and verbs. For example, the object 'the ball' is compatible with a much smaller set of actions (e.g., kicking, throwing, buying) than the subject 'the boy'. Hence, in object-initial transitive sentences, the first constituent carries a lot of information. According to Maurits and colleagues, such sentences are avoided because they violate the *uniform information density* hypothesis, which is a universal tendency for languages to spread information as uniformly as possible across the linguistic signal (Levy, 2005; Jaeger, 2006).

Third, Gibson et al. (2013) argue that a communicative advantage of SVO word order is that the verb separates the subject and the object. The increased distance between subject and object is potentially useful in *semantically reversible* sen-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

tences, such as (2), in which both the subject and object can fulfill the thematic roles of agent and patient.

(2) The boy kicked the girl.

According to Gibson and colleagues, using the verb to separate the subject from the object is useful when the communicative channel is noisy. For example, if the addressee fails to hear the first constituent of (2), they may still infer that 'the girl' is the object, whereas, if the sentence had SOV word order, it would be ambiguous between subject and object.

If certain word orders indeed have cognitive and/or communicative benefits, we may expect that language learners are biased in favour of such word orders when learning a new language, over and above any biases introduced by their native language. In this study, we investigate whether there are such differences in learnability using an *artificial language learning task*. In particular, we hypothesise that there is a correspondence between the ease of learning a language with a particular word order and the typological frequency of this word order in the languages of the world (see Culbertson, 2012; Tily, Frank, & Jaeger, 2011, for similar endeavours).

In the next section, we describe the artificial language learning experiment in more detail. Afterwards, we describe a computational model of joint vocabulary and word order learning to obtain a more direct insight into language learners' prior expectations about word order.

Experiment

In our experiment, participants had to learn an artificial language called Aclapa. The vocabulary of Aclapa comprised seven monosyllabic words that were randomly paired with seven meanings: four entities (square, triangle, circle, heart) and three actions (punching, greeting, photographing).

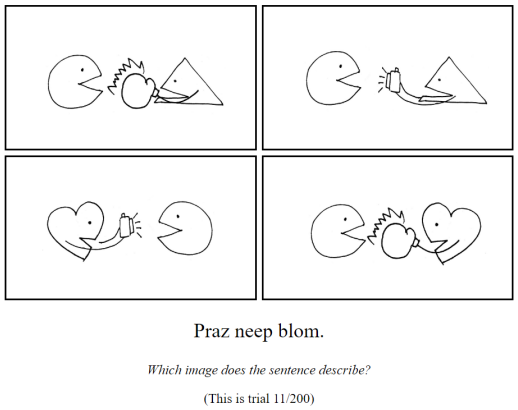

Figure 1: Example trial from the experiment.

The words in Aclapa were used to form three-word transitive sentences. For example, the sentence in Fig. 1 might express the proposition that the triangle punches the circle, as illustrated in the upper left corner. Importantly, the order of subject, verb, and object was varied between participants across all six possible word orders.

Participants learned Aclapa by means of a picture selection task. On each trial, they saw a sentence in Aclapa and four pictures (Fig. 1). They had to click on the picture that matched the meaning of the sentence. Afterwards, they received feedback on whether their response was correct, and, if not, what the correct response was. In this way, participants gradually learned Aclapa by trial-and-error.

To infer participants' prior expectations about the word order of Aclapa, we statistically compared the average number of correct responses for each word order.[1] Here, we assume that, if participants are biased towards certain word orders, they will learn those word orders more easily than others, which will result in a higher number of correct responses.

Obviously, participants will have a strong bias in favour of the word order in their native language. Since we tested native English speakers, participants are expected to have a strong bias for SVO, which is the typical word order in English transitive sentences. Consequently, our primary interest is in whether there are differences in learnability between the remaining five word orders. Recall that, typologically, VOS, OSV, and OVS are highly infrequent. Hence, we may expect that participants have more difficulties learning languages with those word orders when compared to the more commonplace word orders SOV and, to a lesser extent, VSO.

Following this traditional statistical analysis, we seek to obtain a more direct insight into participants' prior expectations about the word order of Aclapa using computational modelling. Thus, we will define a computational model of joint vocabulary and word order learning, based on earlier work by Maurits, Perfors, and Navarro (2009), which we use to obtain a quantitative estimate of participants' pre-experiment initial weights (or "priors") over word orders.

Methods

Participants

386 participants were recruited on Prolific (mean age: 34, standard deviation: 11, range: 18–59; 221 female, 159 male, and 6 other). Participants were randomly assigned to one of the six possible word orders (SVO, SOV, VSO, VOS, OSV, or OVS). There were thus 64 participants in each word order condition, except for SVO, which had 66 participants because of a technical issue.

Materials and procedure

The vocabulary of Aclapa consisted of seven words with seven meanings. The words were taken from the artificial language Brocanto2 (Morgan-Short, Steinhauer, Sanz, & Ullman, 2012) and were simple monosyllabic strings: 'teck',

[1] As clarified below, the participants are not modelled as Bayesian reasoner. Therefore, by 'prior' we will mean the initial expectation over word orders in the context of the belief updating mechanism described below, rather than the initial probabilities in the context of Bayesian update.

'neep', 'blom', 'vode', 'klin', 'praz', and 'yabe'. The possible meanings comprised four types of entities (square, triangle, circle, heart) and three types of actions (punching, greeting, photographing). For each participant, a vocabulary was generated by randomly pairing words with meanings.

Based on the vocabulary and word order, 200 three-word transitive sentences were randomly generated, with the constraint that the subject was never the same as the object. Each of these sentences was paired with four images. The first image depicted the correct meaning of the sentence. The second image depicted the same agent and patient as the first, but showed a different action. The third and fourth images depicted at least one different entity (the agent, the patient, or both). The third image depicted the same action as the first image; the fourth image the same action as the second image. In this way, participants could not deduce the correct image on the basis of the configuration of the images. The four images were shown in a 2×2 grid in random order.

Images randomly showed either the agent on the left and the patient on the right, or the patient on the left and the agent on the right. In this way, the images were unbiased towards either subject-first or object-first word orders.

In the instructions, participants were introduced to the four types of entities and three types of actions. In addition, they were explicitly instructed that the order in which the entities were ordered in the images (i.e., agent on the left and patient on the right, or vice versa) did not influence how these images were described linguistically.

Each trial showed a sentence with four pictures, as described above. Participants were instructed to click on the picture that matched the meaning of the sentence. Participants received feedback after every trial. If they had answered incorrectly, they were shown what the correct answer was. The duration of the feedback ranged from 500 milliseconds to 4 seconds, depending on the stage of the experiment (feedback duration was longer during the first 25 trials to facilitate learning) and the number of correct responses (feedback duration became incrementally shorter after one and four correct responses in a row). In this way, participants were encouraged to do their best to learn Aclapa to speed up the experiment.

The experiment was hosted online on the now-defunct Ibex Farm. A port of the experiment can be accessed via the PCIbex Farm (Zehr & Schwarz, 2018) using the following link: `https://farm.pcibex.net/p/ynpTam/`.

Data treatment

For our analysis, we focus on participants whose performance improved during the experiment. Concretely, we determined, for each participant, whether their performance in the second half of the experiment was significantly above chance (i.e., 25%). For this analysis, we used a one-sample Z-test. Participants with a Z-value below 1.64, corresponding to $p = .1$, were removed from the analysis.

In total, 61 participants were removed. These participants included both participants for whom the experiment was too difficult, and participants who did not seriously engage with the experiment. The number of removed participants did not significantly differ across word orders (SVO: 7, SOV: 13, VSO: 10, VOS: 15, OVS: 6, OSV: 10), as shown by a chi-squared test ($\chi^2(5) = 7.1$, $p = .21$). Hence, there were no clear effects of word order biases on the number of non-learners.

In addition, one participant was removed because they indicated that their native language was not English.

Results

Fig. 2 shows the number of correct responses in each of the word order conditions. The mean number of correct responses was 139 out of 200 (range: 54–194, standard deviation: 39). Comparing the different word order conditions, the mean number of correct responses was the highest for SVO (161), and the lowest for OSV (129), VOS (129), and OVS (131). The means for SOV (143) and VSO (140) were in between these extremes.

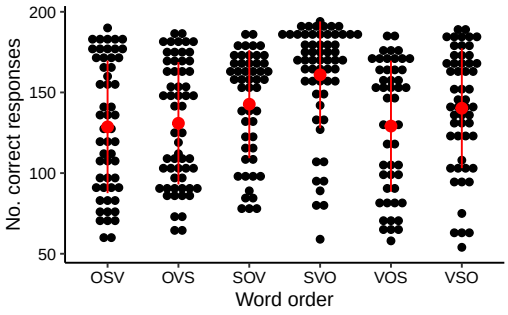

Figure 2: Dotplot showing number of correct responses in each word order condition. The red dot indicates the mean; the red line the standard deviation.

To analyse whether these means were statistically different, we constructed a Poisson generalised linear regression model predicting the number of correct responses based on word order. This analysis was implemented with the 'glm()' function in R (R Core Team, 2023). Post-hoc pairwise comparisons were carried out using the 'glht()' function from the 'multcomp' package, which implements Tukey's method (Hothorn, Bretz, & Westfall, 2008).

Pairwise comparisons indicated that the mean number of correct responses was significantly higher for SVO than for all other word orders (all p's $< .001$). This observation confirms our conjecture that participants have a strong bias in favour of the canonical word order in their native language. In addition, the mean number of correct responses was significantly higher for SOV and VSO than for VOS, OSV, and OVS (all p's $< .001$). None of the other comparisons were statistically significant (all p's $> .89$).

In summary, the results of the experiment suggest the following preference ordering: SVO $>$ {SOV, VSO} $>$ {VOS, OSV, OVS}. Recall that VOS, OSV, and OVS were also typologically the least frequent word orders. Hence, our results suggest that there is a broad correspondence between the

learnability of word orders and their typological frequency, which, in turn, is in line with the idea that there are cognitive and/or communicative motivations underlying the uneven distribution of word orders in the languages of the world.

In this analysis, we took the number of correct responses as a proxy for participants' prior expectations about the word order of Aclapa. In the next section, we use computational modelling to try and obtain a more direct insight into participants' prior expectations.

Cognitive model of language learning

In this section, we propose a computational model of participants' learning process and their behaviour in the experiment, which we use below for model-based statistical analysis of the experimental data. This model takes inspiration from an earlier model of joint vocabulary and word order learning put forward by Maurits et al. (2010). However, our model substantially differs from that earlier model both conceptually and in terms of implementation.

At each timestep, the input to the model comprises, on the one hand, an utterance U consisting of three words u_1, u_2, and u_3, and, on the other hand, three distractor images and a target image T, which consists of three elements: an agent t_A, an action/event t_E, and a patient t_P. The model infers a distribution over languages based on these data, and guesses the target image. The model thus has access to the same information as participants in the experiment; in particular, the model does not have access to the true word order, unlike the linear model reported above.

The model has two components. First, it contains a model of the *agent's beliefs about the language* and how they are updated in response to new observations. Second, it contains a model of *image selection*, which determines the probability that participants select each image given the utterance and the current beliefs about the language. In what follows, we describe these two components in turn.

Beliefs about the language

Each possible language $\mathcal{L}$ can be fully specified by two components. First, an interpretation function I, which associates each word with exactly one of the seven possible meanings (four entities and three actions/events). Since we assume unambiguous vocabularies, there are $(7! =)5{,}040$ possible interpretation functions. Second, a word order ω of the six possible word orders. Since the two components are independent, there is a total of $(5{,}040 \times 6 =)$ 30,240 possible languages.

Participants maintain a distribution over the possible languages. In our model, they do so by storing weights that encode the relative probability of each language. One option would be to store weights for each possible I. However, in order to make the computations feasible, we assume a simpler representation that only encodes association strengths between words and meanings, which we write as '(word, referent)'. This representation requires only $(7 \times 7 =)49$ weights. In addition, we assume that participants keep six weights for the six possible word orders. Taken together, this implies that, at any timestep, participants need to keep track of $(49 + 6 =)55$ weights, which jointly encode a distribution over the possible languages.

Learning is modelled as an update of these weights upon observing the target scene. On each trial, the vocabulary weights are updated as follows. First, for each word u in the sentence and element t in the target image T, we sum the weights of the word orders in which u refers to t. For instance, if u_1 is 'praz' and t_A is a triangle, the value for ('praz', triangle) is the sum of the weights for SVO and SOV, since these are the word orders in which the subject (referring to the agent) is in sentence-initial position. Then, we multiply this quantity by a learning rate parameter γ, add it to the current association strength for all referents in T and words in U, and renormalise the association strengths to obtain, for each signal, a distribution over meanings.

On each trial, the word order weights are updated as follows. First, we compute the total weight of the lexical associations that are compatible with the observation (i.e., the utterance and target image) for each word order. For instance, for word order SVO, this is the product of the association strengths of (u_1, t_A), (u_2, t_E), and (u_3, t_P), which intuitively corresponds to the joint probability of the meanings of the three words in the observation assuming SVO. Then, we multiply this by γ, add it to the respective previous word order weights, and renormalise the word order weights vector.

Image selection

On each trial, given the observed utterance U the agent chooses an image I from four available images, based on their current beliefs encoded in the weights. Call $r_{(\omega, I, u_i)}$ the referent within an image I that word order ω assigns to word u_i. For example, in Figure 1, $r_{(\text{SVO}, \text{topleft}, u_1)} = \text{triangle}$. The probability of choosing each image I given U is computed as follows. First, for each word order ω we compute $(u_1, r_{(\omega, I, u_1)}) \times (u_2, r_{(\omega, I, u_2)}) \times (u_3, r_{(\omega, I, u_3)})$. Then, we multiply each of these by the weight of the corresponding word order and we sum the resulting value. This provides a measure of the relative probability that U refers to each image given the current weights. We use resulting vector as weight parameter for a softmax distribution with temperature parameter α, and sample an image from it.

Model-based statistical analysis

Since the goal of the statistical analysis is to recover the participants' initial weights at the start of the experiment, we leave the prior preferences over word orders as a variable to be estimated from the data. We assume a uniform prior for the interpretation functions, corresponding to the assumption that participants do not associate any words with particular meanings at the beginning of the experiment.

The statistical model has to find a distribution for the variables defining participant behaviour. First, the starting weights for the six possible word orders, representing participants' prior word order biases. Second, the learning weight γ.

Third, the α parameter encoding the noisiness in participants' responses given their underlying beliefs.

We fit three models to the data. First, a *hierarchical model* with partial pooling for participants' word order prior preferences. In this model, there is a population-level variable $\Gamma \sim \text{Gamma}(\alpha = 5, \beta = 2)$, which informs the participants' prior expectations over word order probabilities $\rho \sim \text{Dirichlet}(\alpha = \Gamma)$. Second, a *uniform prior model* with uniform word order prior preferences across participants. Third, a *pooled model* with completely pooled word order prior preferences, where all participants have the same preference parameter with prior distribution $\text{Dirichlet}(\alpha = 1)$. Across all three models, γ was unpooled with a $\text{Gamma}(\alpha = 8, \beta = 15)$ distribution for each participant, and α was unpooled with an $\text{Exponential}(\lambda = 0.5)$ distribution for each participant.

For all models, the information about participants' parameters comes from their choice of scene on each trial. Since the decision for trial i depends only on observations for trials $j < i$, the model needs to keep track of the distribution over languages of each participant on each trial. In other words, to predict participants' behaviour we need to keep track of the whole learning trajectory over the experiment.

In order to approximate the posterior distribution over the unknown parameters, we used HMC sampling with the software package PyMC(v.5). For the hierarchical model and the uniform prior model, we took 1000 samples (with a 1000 samples tuning phase) from 4 chains, for a total of 4000 samples. Unfortunately, the traces show signs of non-convergence for the hierarchical model. Hence, the results should be interpreted with caution.[2] For the pooled model, we took 500 samples (with a 1000 samples tuning phase) from 16 chains, for a total of 8000 samples.

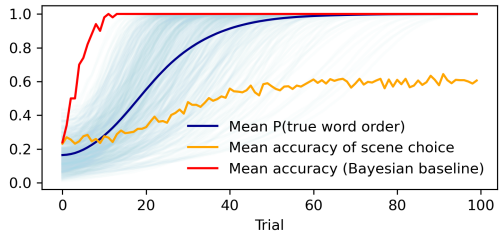

Figure 3: Prior predictive checks for associative model with 100 trials and 500 participants. Light blue lines display the probability of the true word order for each simulated participant. Red line displays a Bayesian baseline.

Fig. 3 shows prior predictive samples for the hierarchical model for a single simulated experiment with 500 participants and 100 trials. As the blue line shows, participants are able to recover the true word order, but there is variation in how fast they do so depending on their initial prior for the true word order and their learning weight. We can therefore expect the statistical model to attribute the variability in accuracy between participants to their prior preferences, as well as their learning rate γ and response noisiness α.

The yellow line in Fig. 3 shows how participants' knowledge is reflected in their task performance. Even participants who correctly understood the language can make mistakes because of the production noise parameter. When participants behave consistently with their predicted knowledge given their observations and learning weight, their noise parameter will have a low estimate, and therefore their behaviour is assumed to be a stronger signal of the other parameters (e.g., the word order priors). The prior predictions of our model are contrasted with a simulated experiment with 50 perfect Bayesian learners (red line). In this case, participants acquire the true language very quickly, and the variation across participants is more limited.

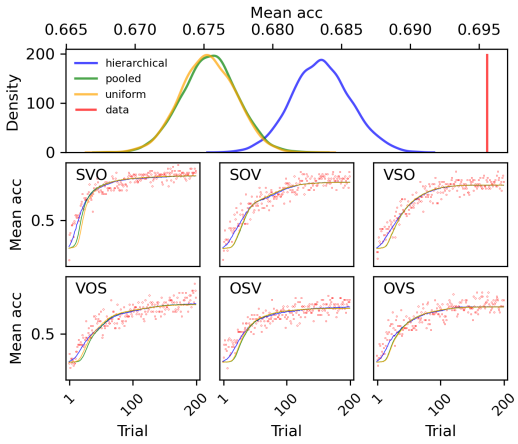

Figure 4: Posterior predictive checks for associative model (8000 simulated experiments). The top plot shows the distribution of average accuracy across participants and trials for the posterior predictive samples from the three models, as well as the real experimental data. The bottom plots show the mean accuracy by trial and word order condition for the three models and the real data.

Fig. 4 shows posterior predictive checks for the average accuracy of all participants (top plot) and split by word order conditions (bottom plots). While posterior samples from the hierarchical model resemble the accuracy of the real participants more closely, all models substantially underestimate the mean accuracy of real participants (red line).

The bottom plots of Fig. 4 reveal where the models fail. While all models capture the main pattern of increasing mean accuracy in the first half of the experiment and subsequent stabilization at less-than-perfect accuracy, they underestimate mean participant accuracy for the early trials, across all word order conditions. This could be because of the factorisation of the language space used in the models or from noisy participants having excessive influence in the posterior estimate. The bottom plots also show that the hierarchical model does the best job at fitting the early trial accuracy levels.

[2] It is not practically possible to run more extensive sampling; the runs we present here took 1000 hours for each chain.

While the hierarchical model failed to converge, the pooled model gives us an estimate of the posterior over word orders that can be inferred from the data assuming the learning model we described above. Specifically, Table 1 reports the posterior probability that each word order has a greater prior probability than each other word order for participants at the beginning of the experiment.

The posterior is uncertain of most comparisons, except for three cases. First, SVO has a credibly higher prior than most other orders. This is the clearest signal from the data, and is in line with what we expect given that we tested native English speakers. Second, and more surprisingly, both SOV and VSO are credibly lower than OVS, and, third, VOS is credibly lower than OSV.

	SVO	SOV	VSO	VOS	OSV
SOV	0.837				
VSO	0.821	0.321			
VOS	1.000	0.466	0.485		
OSV	1.000	0.405	0.422	0.041	
OVS	0.759	0.013	0.035	0.443	0.507

Table 1: Comparison of posterior probability that participants at the beginning of the experiment assigned a greater prior probability to the column word order than the row word order.

The results of the pooled model prima facie contradict the typological data and the linear regression results discussed earlier. While the latter results were broadly aligned with the typological data (i.e., a preference ordering of SVO $>$ {SOV, VSO} $>$ {VOS, OVS, OSV}), the results of the computational modelling exercise are less clear cut, including a surprising higher estimated prior for OVS than SOV and VSO.

In the next section, we offer an explanation for this discrepancy. However, it is also important to emphasise that the modelling results should be interpreted with caution. First, the posterior predictive checks tell us that the model is failing to capture some crucial aspect of participants' learning curve. Second, the hierarchical model, which best captured participants' performance, did not converge despite substantial effort. Since the participants' starting word order priors only subtly show up in the data, it is likely that a better cognitive model is needed to make reliable inferences. Nonetheless, since our model could capture at least some important features of participants' behaviour in the task, we believe it provides a promising first approximation.

General discussion

The languages of the world vary in the way they typically order subject, verb, and object in transitive sentences. However, the frequency of the possible word orders is extremely skewed: most languages have SVO or SOV word order, a few have VSO, and almost none have VOS, OSV, or OVS.

Here, we used an artificial language learning paradigm to experimentally investigate whether languages with less frequent word orders are also more difficult to learn, which would support the idea that there are cognitive and/or communicative biases for certain word orders.

Looking at participants' accuracy throughout the experiment, we confirmed that there was a broad correspondence between learnability and typological frequency: SVO had the highest accuracy, followed by SOV and VSO, followed by VOS, OSV, and OVS.

We further analysed the results using a computational model of language learning to gauge participants' prior expectations of the word order more directly. The results of the latter analysis were more difficult to interpret: although we confirmed the bias for SVO, we also observed a bias for OVS over SOV and VSO.

Although there are important technical reasons for being cautious of the latter finding, we briefly consider a possible reason for why the model inferred a bias towards OVS. Strategically, participants might learn the artificial language by first concentrating on learning the verbs. OVS mirrors the SVO order of participants' native language in having the verb in second position. Once the verb is identified, participants' accuracy increases to 50%, since there are always only two images that display the correct action. Indeed, it may be the case that some participants did not go beyond this stage, and thus failed to learn the meanings of the other words. If so, their behaviour is ambiguous between a bias for SVO and OVS. We aim to explore this possibility in future work.

In our analysis, we considered only simple effects of native language, such that participants find it easier to learn languages with the same word order as their native language. However, it is conceivable that the native language bias is more complex, such that, e.g., people also find it easier to learn languages with a word order that is *similar* to their native language, according to some suitable metric of similarity (e.g., the number of arguments that are in the same position). In order to investigate this possibility, it may be necessary to test participants with a native language that is not SVO. This would also allow us to distinguish the influence of L1 from universal preferences for certain word orders.

In sum, we have provided tentative evidence for an important parallel between learnability and typological frequency, which provides further evidence that there are deep-lying cognitive and/or communicative reasons for universal structural tendencies in the languages of the world.

Acknowledgements

We received four incredibly thorough and thoughtful reviews. We thank the reviewers for their constructive feedback. For reasons of time, we were only able to address some of the issues that were raised; however, we will take all of their valuable feedback on board in future work on this project.

This project was financially supported by Dutch Science Organisation Gravitation Grant "Language in Interaction" 024.001.006.

References

Comrie, B. (1989). *Language universals and linguistic typology*. MIT Press.

Culbertson, J. (2012). Typological universals as reflections of biased learning: Evidencce from artificial language learning. *Language and Linguistics Compass*, *6*, 310–329.

Dryer, M. S. (2013). Order of subject, object and verb (v2020.3). In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Zenodo.

Futrell, R., Hickey, T., Lee, A., Lim, E., Luchkina, E., & Gibson, E. (2015). Cross-linguistic gestures reflect typological universals: A subject-initial, verb-final bias in speakers of diverse languages. *Cognition*, *136*, 215–221.

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, *24*, 1079–1088.

Givón, T. (1979). *On understanding grammar*. Academic Press.

Goldin-Meadow, S., So, W. C., Özyürek, A., & Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences*, *105*, 9163–9168.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), *Universals of language.* MIT Press.

Haider, H. (2023). *"OVS" — A misnomer for SVO languages with ergative alignment.* Retrieved from `https://ling.auf.net/lingbuzz/005680`

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*, 346–363.

Jaeger, T. F. (2006). *Redundancy and syntactic reduction in spontaneous speech*. Doctoral dissertation, Stanford University.

Kemmerer, D. (2012). The cross-linguistic prevalence of SOV and SVO word orders reflects the sequential and hierarchical representation of action in Broca's area. *Language and Linguistics Compass*, *6*, 50–66.

Levy, R. (2005). *Probabilistic models of word order and syntactic discontinuity*. Doctoral dissertation, Stanford University.

Maurits, L., & Griffiths, T. (2014). Tracing the roots of syntax with Bayesian phylogenetics. *Proceedings of the National Academy of Sciences*, *111*, 13576-13581.

Maurits, L., Perfors, A., & Navarro, D. (2009). Joint acquisition of word order and word reference. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th annual conference of the Cognitive Science Society* (pp. 1728–1733). Cognitive Science Society.

Maurits, L., Perfors, A., & Navarro, D. (2010). Why are some word orders more common than others? A uniform information density account. *Advances in Neural Information Processing Systems*, *23*, 1585–1593.

Morgan-Short, K., Steinhauer, K., Sanz, C., & Ullman, M. T. (2012). Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *Journal of Cognitive Neuroscience*, *24*, 933–947.

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Sandler, W., Meir, I., Padden, C., & Aronoff, M. (2005). The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences*, *102*, 2661–2665.

Schouwstra, M., & de Swart, H. (2014). The semantic origins of word order. *Cognition*, *131*, 431–436.

Tily, H., Frank, M., & Jaeger, F. (2011). The learnability of constructed languages reflects typological patterns. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd annual conference of the Cognitive Science Society* (pp. 1364–1369). Cognitive Science Society.

Zehr, J., & Schwarz, F. (2018). *PennController for internet based experiments (IBEX).* Retrieved from `https://doi.org/10.17605/OSF.IO/MD832`