The learnability of constructed languages reflects typological patterns

Harry J. Tily (hjt@mit.edu)
MIT Brain & Cognitive Sciences, 43 Vassar Street, Building 46, Room 3037
Cambridge, MA 02139 USA

Michael C. Frank (mcfrank@stanford.edu)
Stanford University Department of Psychology, 450 Serra Mall
Stanford, CA 94301 USA

T. Florian Jaeger (tiflo@bcs.rochester.edu)
University of Rochester Brain & Cognitive Sciences, Meliora Hall, Box 270268
Rochester, NY 14627 USA

Abstract

A small number of the logically possible word order configurations account for a large proportion of actual human languages. To explain this distribution, typologists often invoke principles of human cognition which might make certain orders easier or harder to learn or use. We present a novel method for carrying out very large scale artificial language learning tasks over the internet, which allows us to test large batteries of systematically designed languages for differential learnability. An exploratory study of the learnability of all possible configurations of subject, verb, and object finds that the two most frequent orders in human languages are the most easily learned, and yields suggestive evidence compatible with other typological and psycholinguistic observations.

Keywords: artificial grammar; language acquisition; language typology; psycholinguistics; word order

Language typology and cognitive universals

Although human languages could in principle place their meaningful elements in any conceivable order, certain patterns crop up much more frequently than would be expected by chance even in unrelated languages. Just taking the order of the three principal sentence components subject, verb and object, there are six possible orders, but the dominant order in over 85% of the world's languages is either SOV or SVO. The table below shows the typological distribution in a sample of 402 languages surveyed by Tomlin (1986).[1]

SOV	SVO	VSO	VOS	OVS	OSV	
45	42	9	3	1	0	(%)

Researchers in language typology have often hypothesized that properties of the human cognitive system could account for these patterns. For example, the fact that the three most frequent word orders are SOV, SVO and VSO can be explained by the cognitive prominence of agentive, animate, or topical referents, and therefore a bias to mention subjects first.

However, comparatively little work has directly used behavioral results to test or build on these typological theories (Jaeger & Tily, 2011, for discussion). This is partly because languages differ in many respects, so any difference in a behavioral measure taken from speakers of two languages might have any number of underlying causes. Probably the most productive behavioral research has been enabled by an observation Hawkins (2004) calls the *Performance-Grammar Correspondence Hypothesis*. Roughly, this holds that the same cognitive principles influence both speakers' choices between variant forms within a language, and the processes of language change which lead a language to hard-code a certain order into its grammar. This allows us to link behavioral effects within a language to cross-linguistic typological trends. For instance, Hawkins shows that English speakers' choices between multiple possible orders such as those in (1) can be explained by a preference to keep as close as possible the verb and the head word of each of its arguments (here the prepositions shown in bold).[2]

(1) a. the woman **waited** [**for** her son] [**in** the cold but not unpleasant wind]

b. the woman **waited** [**in** the cold but not unpleasant wind] [**for** her son]

A similar effect is well documented in behavioral studies of comprehension difficulty (Gibson, 2000). Hawkins' principle can explain Greenberg's (1963) observation that VO languages tend to have prepositions while OV languages have postpositions. The opposite order would result in longer distances between the verb and adpositions, on average.

Similarly, work by Branigan and colleagues (e.g. Branigan et al., 2008) has shown that in languages like Greek and Japanese with variable SO/OS order, participants in memory tasks often "correct" the OS sentences they previously heard to SO order when recalling them. The greater "conceptual accessibility" of agents results in a preference for SO order for speakers of languages with both SO and OS, and may be the reason for the infrequency of primarily OS languages discussed previously.

An alternative way to apply behavioral data to explaining typological distributions is to eschew natural languages,

We thank Masha Fedzechkina for technical assistance, and Cassandra Jacobs and Andrew Wood for their work creating stimuli.

[1] More recent work has suggested that language counts may skew the true distribution by not accounting for relatedness of languages (e.g. Dryer 2009), but this approximate ranking is uncontroversial.

[2] In some cases, words prior to the head are the relevant ones for Hawkins' theory, but this detail does not apply in our examples.

and teach participants constructed languages with the desired properties. Recent work in the artificial grammar paradigm has begun to investigate whether typological patterns can be reproduced in patterns of learnability (see Christiansen 2000; Finley & Badecker 1998; Culbertson & Smolensky 2011). Fedzechkina, Jaeger and Newport (2011) have additionally started to look for a *communicative* explanation for such biases, showing greater preservation of case systems in artificial languages which would otherwise contain ambiguity.

We present results using an artificial grammar paradigm to explore the most widely discussed typological finding, the distribution of dominant basic word orders. Using artificial language lets us test all six possible orders on an equal footing. We introduce a novel web-based paradigm that allows us to conduct experiments with hundreds or potentially thousands of participants in a short time, making possible designs with a large number of conditions. Experiment 1 introduces and tests this new software system, while Experiment 2 presents our first results from a word order experiment using it. Our results suggest that constructed languages with certain basic word orders are learned or used less easily than others.

Experiment 1: Probability matching

Experiment 1 was designed to evaluate a novel methodology for conducting language-learning experiments with large numbers of participants over the web. We use a custom-made Adobe Flash applet developed by the first author (see Figure 1). This applet is capable of displaying audio and video stimuli, and recording the user's responses to various types of test trials. Participants were recruited via Amazon's Mechanical Turk (`http://mturk.com`), an online service where users are paid to perform simple tasks.

We attempted to replicate the well-known phenomenon of *probability matching*, the tendency for adult learners given inconsistent input to replicate the frequency with which they observe different forms. We followed an experiment reported by Hudson Kam and Newport (2005) which varied the frequency with which nouns in a constructed language were heard with a "determiner" word. In subsequent production tasks, participants produced those determiners with roughly the same frequency.

In Hudson Kam and Newport's language, each noun was associated with one of two possible determiners, and a second manipulation varied whether those classes were an arbitrary division of nouns or were semantically based (a mass/count division). We varied a similar semantic variable, the *natural gender consistency* of our determiners. In half the conditions, the two noun classes corresponding to our two determiners separated the male characters from the females. In the other conditions, each class contained both males and females.

Design and materials

We tested 7 language types. Six were constructed by crossing the type of determiner (natural gender vs arbitrary) with the frequency of determiner use (one third vs two thirds vs all

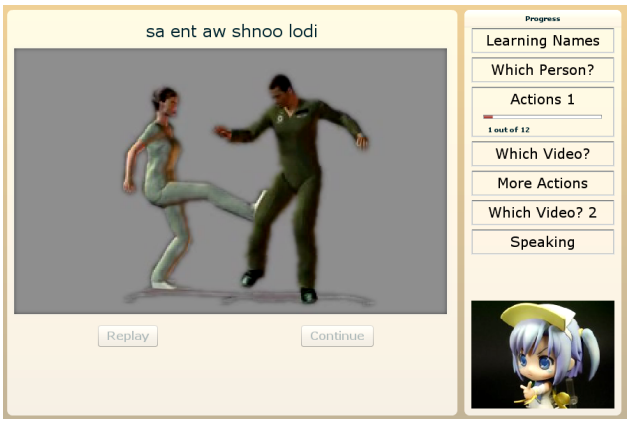

Figure 1: Screenshot during a learning trial *(the nurse kicks the mechanic)*.

nouns). The seventh language had no determiners. Each language contained intransitive sentences, with the word order SV (subject-verb) and transitive, with order SOV (subject-object-verb).[3] Where determiners were present, they preceded the nouns. Thus transitive sentences all followed the template *(det) subject-noun (det) object-noun verb*.

Meanings were supplied by short videos showing human characters performing actions. The videos were rendered in 3D using the EFrontier's Poser software package. There were three female characters (*hippie, nurse, cheerleader*) and three males (*mechanic, conductor, cop*). There were also 2 inanimates (*flower, apple*). In both the arbitrary and gender determiner conditions, one inanimate was assigned randomly to each class. Each animate was rendered performing each of two intransitive actions (*crouch, jump*), and as the agent of each of four transitive actions (*pull, pick up, carry, kick*), with each of the other 7 characters/objects as the patient. This yielded 12 intransitive and 168 transitive videos.

Sounds were recorded as individual words, and sentences were created by concatenation. We created a pool of 6 possible monosyllabic words for use as determiners, and 29 mono- to trisyllabic content words. Each participant saw a different random assignment of words to meanings, by sampling 2 of the 6 possible determiners and 14 of the 29 possible content words for the 8 nouns and 6 verbs.

Procedure

Participants were recruited via Mechanical Turk and paid $0.75 or $1.00 USD. A total of 134 people took part. All were located in the USA and self-reported native English speakers. Participants were randomly assigned to conditions.

Participants received instructions throughout the task from an on-screen cartoon character who told them they would learn an "alien language". In block 1 (noun learning), they saw the 12 intransitive scenes, each accompanied by the

[3] We will use grammmatical function labels S, V & O throughout for simplicity. Strictly, though, S indicates the single argument of an intransitive verb or agent of a transitive, and O the patient.

corresponding sentence in their language via audio and orthographic transcription. After seeing all 12, participants completed 6 forced-choice tasks, where they heard a sentence while watching two previously seen videos of different characters performing the same action played simultaneously side-by-side. Participants clicked on one video to indicate which corresponded to the sentence. Participants who failed to get all 6 correct received feedback and were returned to the beginning of the experiment to try again.

In block 2 (action learning), participants saw 12 transitive actions, 6 with each of two randomly selected verbs, accompanied by audio and transcription. Afterwards, 4 more forced choice trials were shown using previously unseen scenes with these two verbs. Each pair of videos differed in only their verb, subject, or object. Participants received performance feedback (number of correct trials), but continued regardless of performance.

Block 3 (action learning 2) introduced the remaining 2 transitive verbs in the same way as block 2. Afterwards, there were 8 forced choice trials like the previous ones, with each of the 4 transitive verbs being used as the correct video twice.

In the 1/3 and 2/3 determiner conditions, nouns were chosen at random to be given determiners, but each of the three learning blocks was constrained to contain the appropriate proportion overall.

In block 4 (production), participants were instructed to "speak" the language they had learned. On each of 15 trials, a previously unseen transitive action video was played without any sound. Above the video was a written vocabulary list of words from the language. Participants constructed a sentence by clicking on words one at a time, and could hear their sentence by clicking a "play" button, or click on a "reset" button to start again. They were told to click on a "continue" button when their sentence was an appropriate description of the video. Videos in this block had two animate participants.

Results

Two types of behavioral data were collected: the forced choice discrimination tasks seen after blocks 2 and 3, which were combined in the analysis, and the production trials. Data was analyzed using multilevel logistic regression (Jaeger, 2008) with maximal random effects for participant, verb, subject, and object, except where there were too many subjects with ceiling performance to fit such a model, where we fell back on weighted empirical logit regression over subject means (McCullagh & Nelder, 1989). We excluded the no-determiner condition from these analyses to yield a crossed design. We report p values for chi-square model comparison tests associated with the predictors of interest.

Discrimination performance was high in all conditions. Natural gender determiner classes lead to slightly higher performance (93 vs 87%, $p = .01$) but there is no effect of input frequency (main effect: $p = .91$, interaction:$p = .13$). In the no determiner condition, 91% of trials were correct.

Multiple measures were calculated from the production trials. First, we calculated *vocabulary correctness*: sentences are correct under this measure if they contain the correct verb and the two correct nouns, in any order and with or without any determiners. There was no effect of determiner type ($p = .78$), but correctness differed with input frequency (main effect: $p = .03$, interaction: $p = .11$). This was clearly driven by high performance in the 1/3-gender condition (77%, compared with 47-63% in the other 5 conditions. This pattern is unexpected, and we assume it is due to chance. In the no determiner condition, 39% of trials had the correct words.

Next we calculated *word order correctness*. Sentences are considered correct if their verb, subject and object are in the expected order, SOV. We excluded trials which did not have the correct words. Performance is close to ceiling in all conditions (grand mean 89%), showing that when participants get the words right, they almost always use them in the correct order. There was a significant effect of determiner type ($p < .001$) such that natural gender conditions were correct more than arbitrary conditions (92 vs 85%). There was an effect of input frequency ($p < .001$) with performance increasing with greater input frequency (83, 92 and 94%). There was no significant interaction ($p = .95$). In the no determiner condition, performance was 87%.

Figure 2 shows *determiner presence*. This is the proportion of nouns produced by a participant that are preceded by any determiner. There is an obvious effect of input frequency such that participants who saw more determiners in the input produce proportionally more themselves (main effect: $p < .001$) but no effect of determiner type (main effect: $p = .11$; interaction: $p = .13$).

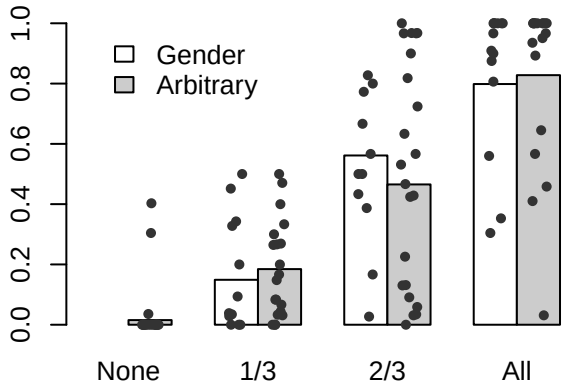

Figure 2: Exp1 proportion of nouns produced with any determiner (bars show condition means, points show subject means).

Finally, Figure 3 shows *determiner correctness*. Each determiner prodcued by a participant is marked correct or incorrect depending on whether it is in the appropriate class for the following noun. There is a main effect of determiner type such that natural gender conditions are more accurate than arbitrary classes (81 vs 69%; $p = .04$) but no effect of input frequency (main effect: $p = .89$; interaction: $p = .73$).

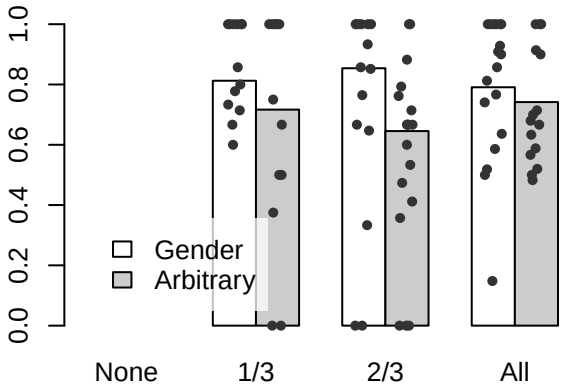

Figure 3: Exp1 proportion of determiners used in correct class.

Discussion

Judging from the discrimination, vocab, and order correctness measures, our participants learn the languages well. They perform close to ceiling in comprehension, and produce fully correct sentences about half the time. Sentences produced with the correct words are almost always in the correct order.

In production, participants produce determiners with frequency depending on their frequency in the input: 19% in the 1/3 of nouns, 58% in the 2/3 of nouns, and 93% in the all nouns conditions. The production frequency of determiners is always slightly lower than the frequency seen in training. This mirrors the results of Hudson Kam and Newport (2005): there, input frequencies of 45, 60, 75 & 100% yielded production frequencies of roughly 30, 45, 65 & 100% respectively.

Hudson Kam and Newport's noun class manipulation had no significant effect on production frequency. Similarly, our manipulation of arbitrary versus natural gender classes has no significant effect on production frequency. However, we found a novel result in that determiner class correctness was greater when the classes corresponded to natural genders. Although the interaction with frequency is not significant, the determiner type effect is smallest in the condition where all nouns in the input had determiners. This might suggest that natural semantic classes are easier to learn with limited exposure but that this effect diminishes with more input.

In conclusion, these results confirm the utility of our web-based language learning method for conducting artificial grammar experiments. Participants were able to learn and use the languages they were taught, and their behavior with respect to inconsistent input closely matches that previously reported for participants learning larger languages over several days in the lab. Extending previous work, we also find evidence that word classes that map onto natural semantic classes are learned better than arbitrary classes.

Experiment 2: Word order variation

Experiment 2 was constructed to test the hypothesis that different basic word orders are observed with different frequencies among the world's languages because some orders are more naturally acquired or used than others. We chose to manipulate the order of the basic constituents S, O and V. Our experiment was intended to determine whether SO languages are more easily learned than OS languages. As discussed earlier, the vast majority of natural languages are SO, and one explanation is a universal cognitive preference to place (salient, agentive) subjects before objects. By testing all 6 possible word orders, we can also check our results for a more general relationship between the learnability of each order and its relative typological frequency.

As a secondary manipulation, we also varied determiner-noun order. There is a typological tendency for VO languages to have determiners which precede the noun, while OV languages more often have determiners which follow the noun (Dryer, 1989).[4] If this correlation emerges from a deeper cognitive property, we might observe differences in the learnability of pre- vs postnominal determiner languages depending on their VO/OV order.

Design and materials

We used exactly the same stimuli as in Experiment 1, but with different languages. This time, determiners appeared with all nouns in all languages, and the two determiner classes were assigned as in the natural gender consistent conditions of Experiment 1. We introduced two new fully crossed manipulations. Each language used one of the six possible basic word orders (SOV, SVO, VSO, OSV, OVS, VOS), and each had either prenominal or postnominal determiners (Det-N vs N-Det). This gives a 6*2 design, of which one cell (SOV/prenominal) was identical to the gender/all nouns condition in Experiment 1.

Procedure

The procedure was identical to that of Experiment 1. 285 participants took part, and were paid $0.75. All were located in the USA and were self-reported native English speakers.

Results

The data was analyzed in the same way as Experiment 1. Figure 4 shows performance on the transitive verb forced choice discrimination trials. Performance on SOV and SVO conditions (94 and 92%) is comparable to the 93% correctness levels seen in natural gender conditions in Experiment 1. However, performance is slightly lower on the other orders VSO, OSV, OVS and VOS (86, 89, 87 and 85% respectively). This word order difference is marginally significant ($p = .06$), and there is a significant effect of determiner order such that prenominal conditions are more accurate (91 vs 87%) but no interaction ($p = .83$).

[4] Real languages differ in whether the various determiner-like functions (e.g. definiteness determination of articles, deictic function of demonstratives) are associated with single or different word categories, and those different functions may be associated with different ordering tendencies. Since our "determiners" are semantically and functionally empty, we necessarily rely on crude typological generalizations.

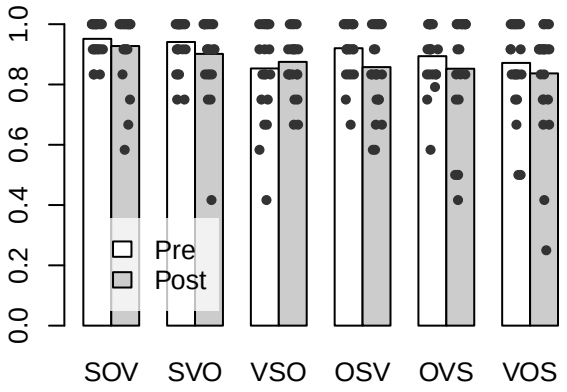

Figure 4: Exp2 proportion of discrimination trials correct (white bars show prenominal, gray postnominal conditions).

Figure 5 shows vocabulary correctness within the production sentences. The highest performance is obtained on the condition with the same word order as English (SVO/prenominal: 59%). Overall correctness is highest in SOV, SVO and OVS (50, 53 and 48%), above VSO, OSV and VOS (34, 45 and 43%). However, neither the 6-way word order variable nor the determiner ordering variable significantly predict vocabulary correctness (word order: $p = .15$; determiner order: $p = .21$; interaction: $p = .12$).

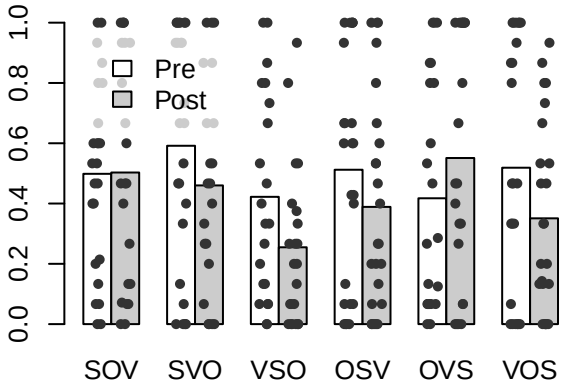

Figure 5: Exp2 proportion of production trials with fully correct words.

Because we were primarily interested in the effect of determiner order with respect to its interaction with VO order, we ran a separate model collapsing together all VO languages and all OV languages. This time, the interaction between word order and determiner order was significant ($p = .02$). This was carried by an advantage for prenominal determiners in VO orders (36 vs 51%) while there was little difference within OV orders (48 vs 47%).

Figure 6 shows word order correctness of the three principal constituents. 100% of trials in the SVO conditions (357 total) were produced in the correct order. SOV shows almost equally high performance (94% overall). Of the remaining orders, VSO (91%) and OVS (88%) are higher than OSV (83%) and VOS (74%). These differences by word order are significant ($p < .01$) although the effect of determiner order is not (main effect: $p = .80$; interaction: $p = .46$). There was also no interaction between determiner order and VO vs OV order ($p = .98$)

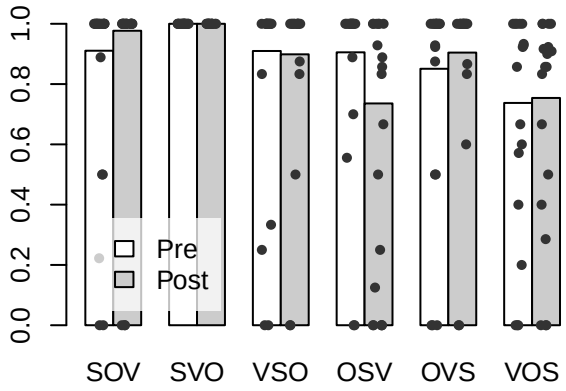

Figure 6: Of Exp2 production trials with correct words, proportion having correct word order.

Discussion

When interpreting the results of Experiment 2, it must be remembered that our participants may have a native language bias towards English-like languages. Indeed, vocabulary correctness is highest in the SVO/prenominal condition, and the only two conditions with 100% accuracy on the order correctness score are the SVO conditions. Order correctness might generally be higher in languages where elements are positioned as in English: the orders where none of the three are in the English position (OSV and VOS) yield the poorest performance. However, there are many other ways a hypothetical native bias could be formulated which do not predict performance. For instance, neither containing English substrings (SV or VO) nor having V-O order are predictive. There is no obvious native language explanation for the high performance on SOV, which is second only to SVO on all both production measures (vocabulary and order) and above SVO in forced choice comprehension. SVO and SOV do have in common a subject which precedes the object, as it does in over 90% of natural languages. This could indicate greater cognitive ease of learning or using languages with SO order. Accordingly, VSO scores third after SVO and SOV in terms of order correctness. In contrast though, in comprehension and in production vocabulary correctness, VSO has the lowest performance of all. Tentatively, then, we suggest that the production task favors SO orders due to a cognitive bias to prefer agents before patients. We do not find any evidence that this bias affects the learning or choice of the words themselves, nor the comprehension of learned languages.

High SOV performance is interesting in light of both the typological frequency of SOV (Dryer (2009) suggests SOV may be twice as frequent as SVO if genera are counted rather than individual languages), and recent work showing SOV advantages. Based on findings from varied paradigms, Goldin-Meadow and colleagues (e.g. Goldin-Meadow, Özyürek &

Mylander, 2008) have argued that SOV is in some sense a cognitively "basic" order for the representation of events: speakers of all languages fall back on SOV in nonlinguistic expression tasks, and emergent sign languages tend to be SOV. Our results can be taken to support this finding, in that participants produce SOV order more accurately than any order besides that of their native language. Moreover, in the comprehension task, SOV performance was in fact slightly above SVO. Thus any SOV advantage is not restricted to production.

A second possible effect that emerges is that of verb mediality. Among the SO orders, there is an SVO advantage on all three measures, which can be explained as native language bias. But among the OS orders too, there is an OVS advantage on the two production measures, and no differences in the comprehension measure. OVS is typologically very rare. However, the psycholinguistic theories of Hawkins and Gibson mentioned in the introduction predict that OVS should be relatively easy to process, since both arguments of the verb are adjacent to it. Gibson, Brink, Piantadosi & Saxe (2011) have suggested that SVO approaches the typological frequency of cognitively basic SOV order because it separates the subject and object, which are similar and therefore confusable. They show that an SOV bias observed in a task with inanimate patients switches to SVO when both agent and patient are animate, as is the case in our materials. Similarly, one explanation for the verb-medial effect here is that adjacent similar noun phrases impede learning or use.

The evidence for determiner order on learnability is mixed. Prenominal determiners led to slightly higher comprehension accuracy, which may simply reflect the familiarity of prenominal determiners to English speaking participants. There was no effect of determiner order on production order correctness, but in vocabulary correctness prenominals led to better performance in VO orders only. It is possible that the main effect of native language obscured potentially more interesting results. Prenominal determiners are favored in the VO conditions, where the native language order and the typological correlation between VO order and determiner order are in alignment. For the OV languages, where those two factors act against each other, there is no difference in correctness. In future research, we hope to explore this question further using pre/postpositions, which have stronger typological tendencies, and with native speakers of OV languages.

Conclusions

We have shown that a web-based language learning task can be used effectively to study the learnability of languages. As expected, English speakers perform well on English-like languages, but they also perform extremely well on SOV languages. This is compatible with previous claims that SO orders — and perhaps SOV specifically — are favored by a default cognitive bias. The internet-based methodology presented here allows an unprecedented amount of data to be collected, allowing for the testing of more and more detailed artificial languages, and using participants with different native language backgrounds. We hope this paradigm will bring new behavioral evidence to the study of language typology.

References

Branigan, H. P., Pickering, M. J., & Tanaka, M. (2008). Contributions of animacy to grammatical function assignment and word order during production. *Lingua*, *118*, 172-189.

Christiansen, M. (2000). *Using artificial language learning to study language evolution: Exploring the emergence of word order universals.* Presented at the Third Conference on the Evolution of Language. Paris, France.

Culbertson, J., & Smolensky., P. (to appear). Testing Greenberg's Universal 18 using an artificial language learning paradigm. In *Proceedings of NELS 40.*

Dryer, M. (1989). Article-noun order. In *CLS 25* (p. 83-97).

Dryer, M. (2009). Problems testing typological correlations with the online WALS. *Ling. Typology*, *13*, 121-135.

Fedzechkina, M., Jaeger, T., & Newport, E. (2011). Functional biases in language learning: Evidence from word order and case-marking interaction. In *Proc. of CogSci 2011.*

Finley, S., & Badecker, W. (2008). Substantive biases for vowel harmony. In *Proceedings of BLS 33.*

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain. Papers from the first Mind Articulation Project Symposium.* Cambridge, MA: MIT Press.

Gibson, E., Brink, K., Piantadosi, S., & Saxe, R. (2011). *Cognitive pressures explain the dominant word orders in language.* Paper presented at CUNY 2011, Stanford, CA.

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* (p. 73-113). London: MIT Press.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford, UK: Oxford University Press.

Hudson Kam, C., & Newport, E. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*, 151-195.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *JML*, *59*(4), 434-446.

Jaeger, T. F., & Tily, H. (2011). On language 'utility': Processing complexity and communicative efficiency. *WIREs: Cognitive Science*, *2*(3), 323-335.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. London, UK: Chapman and Hall.

Tomlin, R. S. (1986). *Basic word order: Functional principles*. London, UK: CroonHelm.