Recursive Inconsistencies Are Hard to Learn: A Connectionist Perspective on Universal Word Order Correlations

Morten H. Christiansen (MORTEN@GIZMO.USC.EDU)
Joseph T. Devlin (JDEVLIN@CS.USC.EDU)
Program in Neural, Informational and Behavioral Sciences
University of Southern California
Los Angeles, CA 90089-2520

Abstract

Across the languages of the world there is a high degree of consistency with respect to the ordering of heads of phrases. Within the generative approach to language these correlational universals have been taken to support the idea of innate linguistic constraints on word order. In contrast, we suggest that the tendency towards word order consistency may emerge from *non-linguistic* constraints on the learning of highly structured temporal sequences, of which human languages are prime examples. First, an analysis of recursive consistency within phrase-structure rules is provided, showing how inconsistency may impede learning. Results are then presented from connectionist simulations involving simple recurrent networks without linguistic biases, demonstrating that recursive inconsistencies directly affect the learnability of a language. Finally, typological language data are presented, suggesting that the word order patterns which are infrequent among the world's languages are the ones which are recursively inconsistent as well as being the patterns which are hard for the nets to learn. We therefore conclude that innate linguistic knowledge may not be necessary to explain word order universals.

Introduction

There is a statistical tendency across human languages to conform to a form in which the head of a phrase consistently is placed in the same position—either first or last—with respect to the remaining clause material. English is considered to be a head-first language, meaning that the head is most frequently placed first in a phrase, as when the verb is placed before the object NP in a transitive VP such as *'eat curry'*. In contrast, speakers of Hindi would say the equivalent of *'curry eat'*, because Hindi is a head-last language. Likewise, head-first languages tend to have *pre*positions before the NP in PPs (such as *'with a fork'*), whereas head-last languages tend to have *post*positions following the NP in PPs (such as *'a fork with'*). Within the Chomskyan approach to language (e.g., Chomsky, 1986) this head direction consistency has been explained in terms of an innate module known as $\overline{X}$-theory which specifies constraints on the phrase structure of languages. It has further been suggested that this module emerged as a product of natural selection (Pinker, 1994). As such, it comes as part of the body of innate linguistic knowledge—i.e., the *Universal Grammar* (UG)—that every child supposedly is born with. All that remains for a child to "learn" about this aspect of her native language is the direction (i.e., head-first or head-last) of the so-called head-parameter.

This paper presents an alternative explanation for word-order consistency based on the suggestion by Christiansen (1994) that language has evolved to fit sequential learning and processing mechanisms existing prior to the appearance of language. These mechanisms presumably also underwent changes after the emergence of language, but the selective pressures are likely to have come not only from language but also from other kinds of complex hierarchical processing, such as the need for increasingly complex manual combination following tool sophistication. On this view, head direction consistency is a by-product of non-linguistic constraints on hierarchically organized temporal sequences. In particular, if recursively consistent combinations of grammatical regularities, such as those found in head-first and head-last languages, are easier to learn (and process) than recursively inconsistent combinations, then it seems plausible that recursively inconsistent languages would simply "die out" (or not come into existence), whereas the recursively consistent languages should proliferate. As a consequence languages incorporating a high degree of recursive inconsistency should be far less frequent among the languages of the world than their more consistent counterparts.

In what follows, we first present an analysis of the structural interactions between phrase structure rules, suggesting that recursive inconsistency results in decreased learnability. The next section describes a collection of simple grammars and makes quantitative learnability predictions based on the rule interaction analysis. The fourth section investigates the learnability question further via connectionist simulations involving networks with a non-linguistic bias towards hierarchical sequence learning. The results demonstrate that these networks find consistent languages easier to learn than inconsistent ones. Finally, typological language data are presented in support of the basic claims of the paper, namely that the word order patterns which are dominant among the world's languages are the ones which are recursively consistent as well as being the patterns which the networks (with their lack of "innate" linguistic knowledge) had the least problems learning.

Learning and Recursive Inconsistency

To support the suggestion that the patterns of word order consistency found in natural language predominately results from non-linguistic constraints on learning, rather than innate lan-

A	→	{a (B)}
B	→	{b A}

Figure 1: A "skeleton" for a set of recursive rules. Curly brackets indicate that the ordering of the constituents can be either as is (i.e., head-first) or in reverse (i.e., head-last), whereas parentheses indicate optional constituents.

guage specific knowledge, it is necessary to point to possible structural limitations emerging from the acquisition process. In the following analysis it is assumed that children only have limited memory and perceptual resources available for the acquisition of their native language. A somewhat similar assumption concerning processing efficiency plays an important role in Hawkins' (1994) performance oriented approach to word order and constituency—although he focuses exclusively on adult processing of language. Although it may be impossible to tease apart the learning-based constraints from those emerging from processing, we hypothesize that basic word order may be most strongly affected by learnability constraints whereas changes in constituency relations (e.g. heavy NP-shifts) may stem from processing limitations.

Why should languages characterized by a mixed set of head-first and head-last rules be more difficult to learn than languages in which all rules are either head-first or head-last? We suggest that the interaction between recursive rules may constitute part of the answer. Consider the "skeleton" for a recursive rule set in Figure 1. From this skeleton four different recursive rule sets can be constructed. These are shown in Figure 2 in conjunction with examples of structures generated from these rule sets. 2(a) and (b) are head-first and head-last rule sets, respectively, and form right and left-branching tree structures. The mixed rule sets, (c) and (d), create more complex tree structures involving center-embeddings. Center-embeddings are difficult to process because constituents cannot be completed immediately, forcing the language processor to keep lexical material in memory until it can be discharged. For the same reason, center-embedded structures are likely to be difficult to acquire because of the distance between the material relevant for the discovery and/or reenforcement of a particular grammatical regularity.

To make the discussion less abstract, we replace "A" with "NP", "a" with "N", "B" with "PP", and "b" with "adp" in Figure 2, and then construct four complex NPs corresponding to the four tree structures:

(1) [$_{NP}$ buildings [$_{PP}$ from [$_{NP}$ cities [$_{PP}$ with [$_{NP}$ smog]]]]]
(2) [$_{NP}$ [$_{PP}$ [$_{NP}$ [$_{PP}$ [$_{NP}$ smog] with] cities] from] buildings]
(3) [$_{NP}$ buildings [$_{PP}$ [$_{NP}$ cities [$_{PP}$ [$_{NP}$ smog] with]] from]]
(4) [$_{NP}$ [$_{PP}$ from [$_{NP}$ [$_{PP}$ with [$_{NP}$ smog]] cities]] buildings]

Notice that in (1) and (2), the prepositions and postpositions, respectively, are always in close proximity to their noun complements. This is not the case for the inconsistently mixed rule sets where all nouns are either stacked up before all the postpositions (3) or after all the prepositions (4). In both cases, the

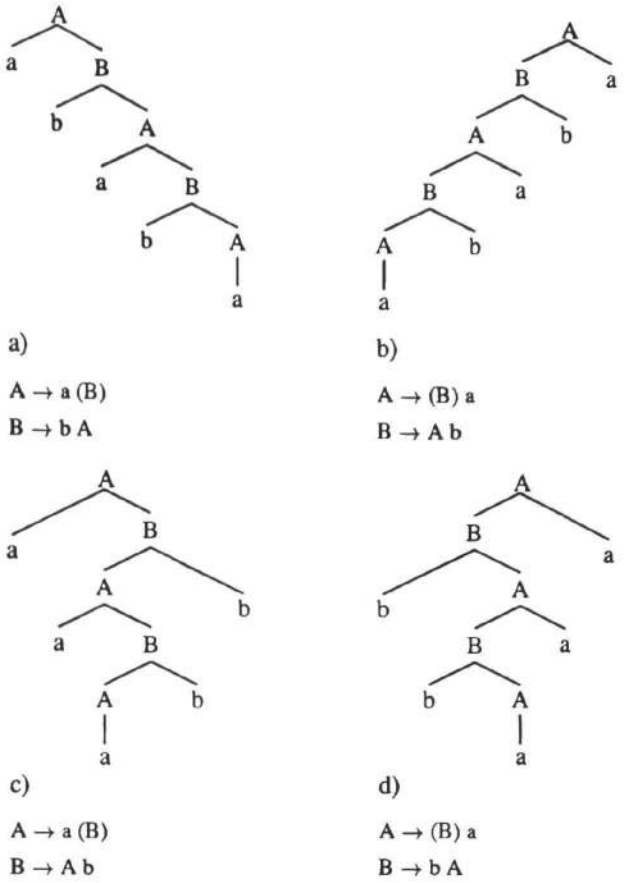

Figure 2: Phrase structure trees built from recursive rule sets that are a) head-first, b) head-last, and c) + d) mixed.

learner has to deduce that "from" and "cities" together form a PP grammatical unit, despite being separated from each other by the PP involving "with" and "smog". This deduction is further complicated by an increase in memory load caused by the latter intervening PP. From a learning perspective, it should therefore be easier to deduce the underlying structure found in (1) and (2) compared with (3) and (4). Given these considerations we define the following learning constraint on recursive rule interaction:

> *Recursive Rule Interaction Constraint (RRIC):* If a set of rules are mutually recursive (in the sense that they each directly call the other(s)) and do not obey head direction consistency, then this rule set will be more difficult to learn than one in which the rules obey head direction consistency.

The RRIC covers rule interactions as exemplified by the skeleton rule set in Figure 1, but leaves out cases where rules do not call each other directly. Figure 3 shows examples of such non-direct rule interactions. For a system which has to learn subject noun/verb agreement, SOV-like languages with structures such as 3(a) are problematic because dependencies generally will be long (and thus more difficult to learn given

memory restrictions). It is moreover not clear to the learner whether *'with delight'* should attach to *'love'* or to *'share'* in *'people in love with delight share'*. In contrast, subject noun/verb agreement should be easier to acquire in SVO languages involving 3(b) since the dependencies will tend to be shorter than in 3(a). Notice also that there is no ambiguity

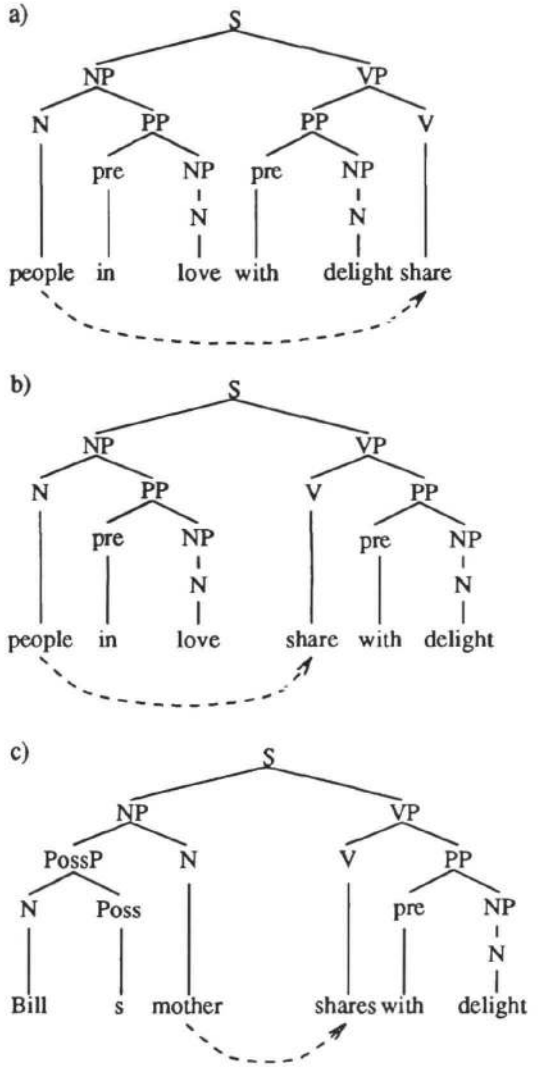

Figure 3: Phrase structure trees for a) an SOV-style language with prepositions, b) an SVO language with prepositions, and c) an SVO language with prepositions and prenominal possessive genitives. The dotted arrows indicate subject noun/verb agreement dependencies.

with respect to the attachment of *'with delight'* in *'people in love share with delight'*[1]. Languages involving constructions such as 3(a) are therefore likely to be harder to learn than

[1]Of course, if we include an object NP then ambiguity may arise as in *'saw the man with the binoculars'*; but this would also be true of SOV-like languages involving 3(a), e.g., *'with the binoculars the man saw'*

S	→	NP VP	
NP	→	{N (PP)}	(1)
PP	→	{adp NP}	(2)
VP	→	{V (NP) (PP)}	(3)
NP	→	{N PossP}	(4)
PossP	→	{Poss NP}	(5)

Figure 4: The grammar "skeleton" used to create the 32 languages for the simulations. Curly brackets indicate that the ordering of the constituents can be either as is (i.e., head-first) or in reverse (i.e., head-last), whereas parentheses indicate optional constituents.

those which include 3(b). Whereas the comparison between 3(a) and (b) indicate a learning motivated preference towards head direction consistency there are exceptions to this trend. One of these exception occurs in English which is predominately head-first, but nevertheless also involves some head-last constructions as exemplified in 3(c). Here the prenominal possessive genitive phrase is head-last whereas the remaining structures are head-first. Interestingly, this inconsistency may facilitate the learning of subject noun/verb agreement since this mix of head-first and head-last structure results in shorter agreement dependencies.

The analysis of rule interactions presented here suggests why certain structures will be more difficult to learn than others. In particular, inconsistency within a set of recursive rules is likely to create learnability problems because of the resulting center-embedded structures, whereas interactions between sets of rules can either impede (as in 3a) or facilitate learning (as in 3c). Of course, other aspects of language (e.g., concord morphology) are also likely to play a part in determining the learnability of a given language, but the analysis above indicates *ceteris paribus* which language structure should be easy to learn and therefore occur more often among the set of human languages. Next, the above analysis is used to make predictions about the difficulty of learning a set of 32 simple grammars.

Grammars and Predictions

In order to test the hypothesis that non-linguistic constraints on acquisition restrict the set of languages that are easily learnable, 32 grammars were constructed for a simulation experiment. Figure 4 shows the grammar skeleton from which these grammars were derived. We have focused on SVO and SOV languages which is why the sentence level rule is not reversible. The numbers on the right hand-side of the remaining five rules refer to the position of a binary variable in a 5-place vector, with the value "1" denoting head-first ordering and "0" head-last. Each of the 32 possible grammars can thus be characterized by a vector, determining the head direction of each of the five rules. The "name" of a grammar is simply the binary number of the vector. For example, the vector "11100" (binary for 28) corresponds to an "English" grammar in which the three first rules are head-first while the rule set capturing possessive genitive phrases (4 and 5) is head-last. Given this naming convention, grammar 0 produces an

all head-last language whereas grammar 31 generates an all head-first language. The remaining grammars 1 through 30 capture languages with differing degrees of head ordering inconsistency.

Given the analysis presented in the previous section we can evaluate each grammar and assign it a number—its *inconsistency penalty*—indicating its degree of recursive inconsistency. The RRIC predicts that inconsistent recursive rule sets should have a negative impact on learning. The grammar skeleton has two possibilities for violating the RRIC: a) the PP recursive rules set (rules 1 and 2), and b) the PossP recursive rule set (rules 4 and 5). Since a PP can occur inside both NPs and VPs, a RRIC violation within this rule set is predicted to impair learning more than a RRIC violation within the PossP recursive rule set. RRIC violations within the PP rule set were therefore assigned an inconsistency penalty of 2, and RRIC violations within the PossP rule set an inconsistency penalty of 1. Consequently, each grammar was assigned an inconsistency penalty ranging from 0 to 3. For example, a grammar which involved RRIC violations of both the PP and the PossP recursive rule sets (e.g., grammar 10110) was assigned a penalty of 3, whereas a grammar with no RRIC violations (e.g., grammar 11100) received a 0 penalty. While other factors are likely to influence the learnability of individual grammars[2], we concentrate on the two RRIC violations to keep the number of free parameters small. In the next section, the inconsistency penalty for a given grammar is used to predict network performance on that grammar.

Simulations

The predictions regarding the learning difficulties associated with recursive inconsistencies are couched in terms of rule interactions. The question remains whether non-symbolic learning devices, such as neural networks, will be sensitive to RRIC violations. The Simple Recurrent Network (SRN) (Elman, 1990) provides a useful tool for the investigation of this question because it has been successfully applied in the modeling of both non-linguistic sequential learning (e.g., Cleeremans, 1993) and language processing (e.g., Christiansen, 1994; Christiansen & Chater, in submission; Elman, 1990, 1991). An SRN is essentially a standard feedforward neural network equipped with an extra layer of so-called context units. The SRN used in all our simulations had 8 input/output units as well as 8 hidden units and 8 context units. At a particular time step t, an input pattern is propagated through the hidden unit layer to the output layer. At the next time step, $t+1$, the activation of the hidden unit layer at time t is copied back to the context layer and paired with the current input. This means that the current state of the hidden units can influence the processing of subsequent inputs, providing a limited ability to deal with integrated sequences of input presented successively. Thus, rather than having a linguistic bias, the SRN is biased towards the learning of hierarchically organized sequential structure.

In the simulations, SRNs were trained to predict the next lexical category in a sentence, using sentences generated by the 32 grammars derived from the grammar skeleton in Figure 4. Each unit in the input/output layers corresponded to one of seven lexical categories or an end of sentence marker: singular/plural noun (N), singular/plural verb (V), singular/plural possessive genitive affix (Poss), and adposition (adp). Although these input/output representations abstract away from many of the complexities facing language learners, they suffice to capture the fundamental aspects of grammar learning important to our hypothesis. By arbitrarily assigning probabilities to each branch point in the skeleton, six corpora of grammatical sentences were randomly generated for each grammar, five training corpora and one test corpus. Each corpus contained 1000 sentences of varying length.

Following successful training, an SRN will tend to output a probability distribution of possible next items given the previous sentential context. For example, if the net trained on the "English" grammar (11100) had received the sequence 'N(sing) V(sing) N(plur)' as input, it would activate the units corresponding to the possessive genitive suffix, Poss(plur), the preposition, adp, and the end of sentence marker. In order to assess how well the nets have learned the grammatical regularities generated by a particular grammar it makes little sense to compare network outputs with their respective targets, say, adp in the above example. Making such a comparison would only allow for an assessment of how well a network has memorized particular sequences of lexical categories. Instead, we assessed network performance in terms of how close the output was to the full conditional probabilities as found in the training corpus. In the above example, the full conditional probabilities would be .105 for Poss(plur), .375 for adp, and .48 for the end of sentence marker. Results are therefore reported in terms of the Mean Squared Error (MSE) between network predictions for the test corpus and the empirically derived full conditional probabilities.

For each of the 32 grammars, we conducted 25 simulations according to a 5×5 set-up, with the five different training corpora and five different initial configurations of the network weights, resulting in a total of $(32\times5\times5)$ 800 network simulations. In these simulations, all other factors remained constant[3]. However, because the sentences in each training corpus were randomly produced, they varied in length. Consequently, to avoid training one net more than another, epochs

[2]For example, the grammars used in the simulations reported below include subject noun/verb agreement. This introduces a bias towards SVO languages because SOV languages will tend to have more lexical material between the subject noun and the verb. In SOV languages case marking are often used to distinguish subjects and objects and this may facilitate learning. For simplicity we have left such considerations out of the current simulations—even though we are aware that they may affect the learnability of particular grammar fragments, and that including them would plausibly improve the fit between our simulations and the typological data.

[3]The *Tlearn* simulator (available from Center for Research on Language, UCSD) was used in all simulations, with identical learning parameters for each net: learning rate: .01; momentum: .95; initial weight randomization: [-.1, .1].

were calculated not in sentences, but in words. In the simulations, 1000 words constituted one epoch of training.

After training each network for 7 epochs, they were tested on the separate test corpus. For each grammar, the average MSE was calculated for the 25 networks. In order to investigate whether the networks were sensitive to violations of the RRIC, a regression analysis was conducted with the inconsistency penalty assigned to each grammar as a predictor of the average network MSE for the 32 grammars. Figure 5 illustrates the result of this analysis, demonstrating a very strong correlation between inconsistency penalty and MSE ($r = .83, F(1,31) = 65.28, p < .0001$)[4]. The higher the inconsistency penalty is for a grammar, the higher the MSE is for the nets trained on that grammar. In order words, the networks are highly sensitive to violations of the RRIC in that increasing recursive inconsistency results in an increase in learning difficulty (measured in terms of MSE). In fact, focusing on PP and PossP violations of the RRIC allows us to account for 68.5% of the variance in MSE.

This is an important result because it is not obvious that the SRNs should be sensitive to inconsistencies at the structural level. Recall that the networks only were presented with lexical categories one at a time, and that structural information about grammatical regularities had to be induced from the way the lexical categories combine in the input. No explicit structural information was provided, yet the networks were sensitive to the structural inconsistencies exemplified by the RRIC violations. In this connection, it is worth noting that Christiansen & Chater (in submission) have shown that increasing the size of the hidden/context layers (beyond a certain minimum) does not affect SRN performance on center-embedded constructions (i.e., structures which are recursively inconsistent structures according to the RRIC). This suggests that the present results may not be dependent on the specific size of the SRNs used here, nor is it likely to depend on the size of the training corpus. Together, these and the present results provide support for the notion that SRNs constitute viable models of natural language processing. Next, this notion is further corroborated by typological language evidence.

Comparisons with Typological Language Data

The present work presupposes that the kinds of structure that the networks find easy to learn should also be the kinds of structure that humans acquire without much effort. Following the suggestion by Christiansen (1994) that only languages that are easy to learn should proliferate, we investigated whether the kinds of structures that the nets found hard to learn were also likely not to be well-represented among the world's languages. The FANAL database developed by Matthew Dryer was used in this investigation. It contains typological information about 625 languages, divided into 252 genera (i.e., groups or families of languages which most typological linguists would consider genetically related; e.g., the group of Germanic languages—see Dryer, 1992, for further details). Unfortunately, the database does not contain the information necessary for a search for all the 32 word order combinations used in the simulations. It was possible to search for partial combinations involving either the PP recursive rule set or the PossP recursive rule set, but only for consistent combinations of these.

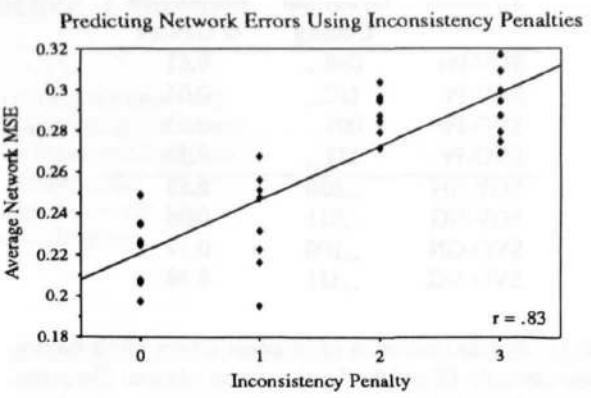

Figure 5: Prediction of the average network MSE for a given grammar using the inconsistency penalty assigned to that grammar.

With respect to the PP recursive rule set we searched for genera which had either SVO or SOV structure and which were either prepositional or postpositional. For the PossP recursive rule set we searched for SVO and SOV languages which had either prenominal or postnominal genitives. Table 1 contains the results from the FANAL search. For each of the two recursive rule sets the proportion of genera incorporating this structure was calculated based on the total number of genera found for that rule set. For example, FANAL found 99 genera with a value for the PP search parameters, such that the SOV-Po proportion of .61 corresponds to 60 genera.

Not surprisingly, SOV genera with postpositions are strongly preferred over SOV genera with prepositions, whereas SVO genera with prepositions are preferred over SVO genera with postpositions. The PossP search shows that there is a strong preference for SOV genera with postnominal genitives over SOV genera with prenominal genitives, but that SVO languages only has a weak preference for prenominal genitives over postnominal genitives. Together the results from the two FANAL searches support our hypothesis that recursive inconsistencies tend to be infrequent among the world's languages.

The results from the FANAL search were interpreted in terms of the 32 grammars, such that a grammar was assigned a number indicating the average proportion of genera for rules

[4] Although the difference in MSE is small (ranging from .1953 to .317), it should be noted that the average standard error of the mean at epoch 7 across all 800 simulations was only .001. Thus, practically all the MSE differences are statistically significant. In addition, when the inconsistency penalties were used as predictors of the average MSE across epoch 1 through 7, a significant correlation ($r = .51, F(1,31) = 10.36, p < .004$) was still obtained—despite the large amount of noise that averaging across 7 epochs produces.

Structure	Grammar Coding	Proportion of Genera
SOV-Po	**000_**	**0.61**
SOV-Pr	110_	0.03
SVO-Po	001_	0.03
SVO-Pr	**111_**	**0.33**
SOV-GN	**_000**	**0.62**
SOV-NG	_011	0.06
SVO-GN	_100	0.12
SVO-NG	**_111**	**0.20**

Table 1: Average proportion of language genera which contain structures from the PP and the PossP recursive rule sets. The grammar codings in bold typeface correspond to consistent rule combinations. The proportions of genera in boldface indicate the preferred combination from a pairwise comparison of two rule combinations (e.g., SOV-GN vs. SOV-NG).

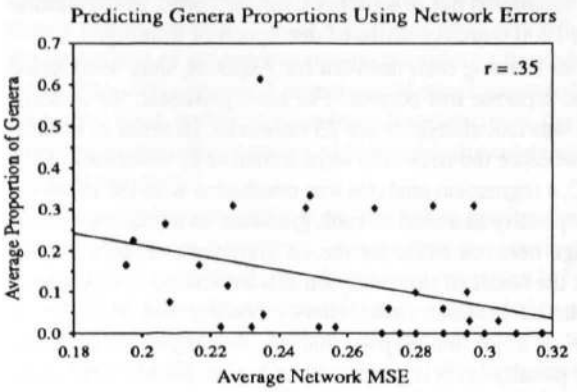

Figure 6: Prediction of the average proportion of genera which contain the particular structures coded for by a grammar using the average network MSE for that grammar.

1-3 (PP search) and rules 3-5 (PossP search). E.g., the PossP combination _000 yielded a proportion of .62 which was assigned to the grammars 00000, 01000, 10000, and 11000. Each of the two FANAL searches covers a set of 16 grammars (with some overlap between the two sets). Grammars with only one proportion value were assigned an additional second value of 0, and grammars with no assigned proportion values were assigned a total value of 0. Finally, the value for each grammar was averaged (e.g., for grammar 00000 the final value was: $(.61 + .62)/2 = .615$).

In Figure 6 the average network MSE for each grammar is used to predict the average proportion of genera that contain the rule combinations coded for by that particular grammar. The figure indicates that the higher the network MSE is for a grammar, the lower the average proportion of genera is for that grammar ($r = .35, F(1, 31) = 4.20, p < .05$). That is, genera involving rule combinations that are hard for the networks to learn tend to be less frequent than genera involving rule combinations that the networks learn more easily (at least for the word order patterns focused on in this paper). The tendency towards recursive consistency among the languages of the world is also confirmed when we use the inconsistency penalties to predict the average proportion of genera for each grammar ($r = .57, F(1, 31) = 14.06, p < .001$).

Conclusion

In this paper, we have provided an analysis of recursive inconsistency and its negative impact on learning, and showed that the SRN—a connectionist learning mechanism with no specific linguistic knowledge—was indeed sensitive to such inconsistencies. A comparison with typological language data revealed that the recursively inconsistent language structures which the SRN had problems learning tended to be infrequent across the world's languages. Together these results suggest that universal word order correlations may emerge from non-linguistic constraints on learning, rather than being a product of innate linguistic knowledge. The broader implication of this suggestion for theories of language acquisition is, if true, that learning may play a bigger role in the acquisition process than typically assumed by proponents of UG. Word order consistency is one of the language universals which have been taken to require innate linguistic knowledge for its explanation. However, we have presented results which challenges this view, and envisage that other so-called linguistic universals may be amenable to explanations which seek to account for the universals in terms of non-linguistic constraints on learning and/or processing.

Acknowledgments

We thank Matthew Dryer for permission to use and advice on using his FANAL database, and Anita Govindjee, Jack Hawkins and Jim Hoeffner for commenting on an earlier version of this paper.

References

Chomsky, N. (1986). *Knowledge of Language*. New York: Praeger.

Christiansen, M.H. (1994). *Infinite Languages, Finite Minds: Connectionism, Learning and Linguistic Structure*. Doctoral dissertation, Centre for Cognitive Science, University of Edinburgh.

Christiansen, M.H. & Chater, N. (in submission). Toward a Connectionist Model of Recursion in Human Linguistic Performance.

Cleeremans, A. (1993). *Mechanisms of Implicit Learning: Connectionist Models of Sequence Processing*. Cambridge, MA: MIT Press.

Dryer, M.S. (1992). The Greenbergian Word Order Correlations. *Language*, **68**, 81–138.

Elman, J.L. (1990). Finding Structure in Time. *Cognitive Science*, **14**, 179–211.

Elman, J.L. (1991). Distributed Representation, Simple Recurrent Networks, and Grammatical Structure. *Machine Learning*, **7**, 195–225.

Hawkins, J.A. (1994). *A Performance Theory of Order and Constituency*. UK: Cambridge University Press.

Pinker, S. (1994). *The Language Instinct: How the Mind Creates Language*. New York: NY: William Morrow and Company.