Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Languages

Shauli Ravfogel[1] **Yoav Goldberg**[1,2] **Tal Linzen**[3]
[1]Computer Science Department, Bar Ilan University
[2]Allen Institute for Artificial Intelligence
[3]Department of Cognitive Science, Johns Hopkins University
{shauli.ravfogel, yoav.goldberg}@gmail.com, tal.linzen@jhu.edu

Abstract

How do typological properties such as word order and morphological case marking affect the ability of neural sequence models to acquire the syntax of a language? Cross-linguistic comparisons of RNNs' syntactic performance (e.g., on subject-verb agreement prediction) are complicated by the fact that any two languages differ in multiple typological properties, as well as by differences in training corpus. We propose a paradigm that addresses these issues: we create synthetic versions of English, which differ from English in one or more typological parameters, and generate corpora for those languages based on a parsed English corpus. We report a series of experiments in which RNNs were trained to predict agreement features for verbs in each of those synthetic languages. Among other findings, (1) performance was higher in subject-verb-object order (as in English) than in subject-object-verb order (as in Japanese), suggesting that RNNs have a recency bias; (2) predicting agreement with both subject and object (polypersonal agreement) improves over predicting each separately, suggesting that underlying syntactic knowledge transfers across the two tasks; and (3) overt morphological case makes agreement prediction significantly easier, regardless of word order.

1 Introduction

The strong performance of recurrent neural networks (RNNs) in applied natural language processing tasks has motivated an array of studies that have investigated their ability to acquire natural language syntax without syntactic annotations; these studies have identified both strengths (Linzen et al., 2016; Giulianelli et al., 2018; Gulordava et al., 2018; Kuncoro et al., 2018; van Schijndel and Linzen, 2018; Wilcox et al., 2018) and limitations (Chowdhury and Zamparelli, 2018; Marvin and Linzen, 2018; Wilcox et al., 2018).

Most of the work so far has focused on English, a language with a specific word order and relatively poor morphology. Do the typological properties of a language affect the ability of RNNs to learn its syntactic regularities? Recent studies suggest that they might. Gulordava et al. (2018) evaluated language models on agreement prediction in English, Russian, Italian and Hebrew, and found worse performance on English than the other languages. In the other direction, a study on agreement prediction in Basque showed substantially *worse* average-case performance than reported for English (Ravfogel et al., 2018).

Existing cross-linguistic comparisons are difficult to interpret, however. Models were inevitably trained on a different corpus for each language. The constructions tested can differ across languages (Gulordava et al., 2018). Perhaps most importantly, any two natural languages differ in a number of typological dimensions, such as morphological richness, word order, or explicit case marking. This paper proposes a controlled experimental paradigm for studying the interaction of the inductive bias of a neural architecture with particular typological properties. Given a parsed corpus for a particular natural language (English, in our experiments), we generate corpora for synthetic languages that differ from the original language in one of more typological parameters (Chomsky, 1981), following Wang and Eisner (2016). In a synthetic version of English with a subject-object-verb order, for example, sentence (1-a) would be transformed into (1-b):

(1) a. The man eats the apples.
b. The man the apples eats.

Proceedings of NAACL-HLT 2019, pages 3532–3542
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

Original		they say the broker took them out for lunch frequently . *(they, broker: subjects; say, took: verbs; them: object)*
Polypersonal agreement		they saykon the broker tookkarker them out for lunch frequently . *(kon: plural subject; kar: singular subject; ker: plural object)*
Word order variation	SVO	they say the broker took out frequently them for lunch .
	SOV	they the broker them took out frequently for lunch say .
	VOS	say took out frequently them the broker for lunch they.
	VSO	say they took out frequently the broker them for lunch .
	OSV	them the broker took out frequently for lunch they say .
	OVS	them took out frequently the broker for lunch say they *(they, broker: subjects; say, took: verbs; them: object)*
Case systems	Unambiguous	theykon saykon the brokerkar tookkarker theyker out for lunch frequently . *(kon: plural subject; kar: singular subject; ker: plural object)*
	Syncretic	theykon saykon the brokerkar tookkarkar theykar out for lunch frequently . *(kon: plural subject; kar: plural object/singular subject)*
	Argument marking	theyker sayker the brokerkin tookkerkin theyker out for lunch frequently . (*ker: plural argument; kin: singular argument*)

Figure 1: The sentences generated in our synthetic languages based on an original English sentence. All verbs in the experiments reported in the paper carried subject and object agreement suffixes as in the polypersonal agreement experiment; we omitted these suffixes from the word order variation examples in the table for ease of reading.

We then train a model to predict the agreement features of the verb; in the present paper, we focus on predicting the plurality of the subject and the object (that is, whether they are singular or plural). The subject plurality prediction problem for (1-b), for example, can be formulated as follows:

(2) The man the apples ⟨singular/plural subject?⟩.

We illustrate the potential of this approach in a series of case studies. We first experiment with polypersonal agreement, in which the verb agrees with both the subject and the object (§3). We then manipulate the order of the subject, the object and the verb (§4), and experiment with overt morphological case (§5). For a preview of our synthetic languages, see Figure 1.

2 Setup

Synthetic Language Generation We used an expert-annotated corpus, to avoid potential confounds between the typological parameters we manipulated and possible parse errors in an automatically parsed corpus. As our starting point, we took the English Penn Treebank (Marcus et al., 1993), converted to the Universal Dependencies scheme (Nivre et al. 2016) using the Stanford converter (Schuster and Manning, 2016). We then manipulated the tree representations of the sentences in the corpus to generate parametrically modified English corpora, varying in case systems, agreement patterns, and order of core elements. For each parametric version of English, we recorded the verb-argument relations within each sentence, and created a labeled dataset. We exposed our models to sentences from which one of the verbs was omitted, and trained them to predict the plurality of the arguments of the unseen verb. The following paragraph describes the process of collecting verb-argument relations; a detailed discussion of the parametric generation process for agreement marking, word order and case marking is given in the corresponding sections. We have made our synthetic language generation code publicly available.[1]

Argument Collection We created a labeled agreement prediction dataset by first collecting verb-arguments relations from the parsed corpus. We collected nouns, proper nouns, pronouns, adjectives, cardinal numbers and relative pronouns connected to a verb (identified by its part-of-speech tag) with an *nsubj*, *nsubjpass* or *dobj* dependency edge, and record the plurality of those arguments. Verbs that were the head of a clausal complement without a subject (*xcomp* dependencies) were excluded. We recorded the plurality of the dependents of the verb regardless of whether the tense and person of the verb condition agreement in English (that is, not only in third-person

[1] https://github.com/Shaul1321/rnn_typology

Prediction task	Subject accuracy	Object accuracy	Object recall
Subject	94.7 ± 0.3	-	-
Object	-	88.9 ± 0.26	81.8 ± 1.4
Joint	95.7 ± 0.23	90.0 ± 0.1	85.4 ± 2.3

Table 1: Results of the polypersonal agreement experiments. "Joint" refers to multitask prediction of subject and object plurality.

	Singular	Plural
Subject	-kar	-kon
Object	-kin	-ker
Indirect Object	-ken	-kre

Table 2: Case suffixes used in the experiments. Verbs are marked by a concatenation of the suffixes of their corresponding arguments.

present-tense verbs). For relative pronouns that function as subjects or objects, we recorded the plurality of their referent; for instance, in the phrase *Treasury bonds, which pay lower interest rates*, we considered the verb *pay* to have a plural subject.

Prediction Task We experimented both with prediction of one of the arguments of the verb (subject or object), and with a joint setting in which the model predicted both arguments of each verb. Consider, for example, the prediction problem (3) (the verb in the original sentence was *gave*):

(3) The state ⟨verb⟩ CenTrust 30 days to sell the Rubens .

In the joint prediction setting the system is expected to make the prediction ⟨subject: singular, object: plural⟩. For each argument, the model predicts one of three categories: SINGULAR, PLURAL or NONE. The NONE label was used in the object prediction task for intransitive verbs, which do not have an object; it was never used in the subject prediction task.

Model We used bidirectional LSTMs with 150 hidden units. The bidirectional LSTM's representation of the left and right contexts of the verb was fed into a multilayer perceptron (MLP) with two hidden layers of sizes 100 and 50. We used independent MLPs to predict subject and object plurality. To capture morphological information, words were represented as the sum of the word embedding and embeddings of the character n-grams that made up the word.[2]

The model (including the embedding layer) was trained end-to-end using the Adam optimizer (Kingma and Ba, 2014). For each of the experiments described in the paper, we trained four models with different random initializations; we report averaged results alongside standard deviations.

3 Polypersonal Agreement

In languages with polypersonal agreement, verbs agree not only with their subject (as in English), but also with their direct object. Consider the following Basque example:[3]

(4)	*Kutxazain-ek*	*bezeroa-ri*	*liburu-ak*
	cashier-PL.ERG	customer-SG.DAT	book-PL.ABS
	eman dizkiote		
	gave they-them-to-her/him		

The cashiers gave the books to the customer.

Information about the grammatical role of certain constituents in the sentence may disambiguate the function of others; most trivially, if a word is the subject of a given verb, it cannot simultaneously be its object. The goal of the present experiment is to determine whether jointly predicting both object and subject plurality improves the overall performance of the model.

Corpus Creation In sentences with multiple verbs, agreement markers on verbs other than the prediction target could plausibly help predict the features on the target verb. In a preliminary experiment, we did not observe clear differences between different verb marking schemes (e.g., avoiding marking agreement on verbs other than the prediction target). We thus opted for full marking in all experiments: verbs are modified with suffixes that encode the number of all their arguments (see Figure 1). The suffixes we used for verbs are a concatenation of the respective case suffixes of their arguments (Table 2). For consistency, we remove plurality markers from English

[2] Specifically, let E_t and E_{ng} be word and n-gram embedding matrices, and let t_w and NG_w be the word and the set of all n-grams of lengths 1 to 5, for a given word w. The final vector representation of w, e_w, is given by $e_w = E_t[t] + \sum_{ng \in NG_w} E_{ng}[ng]$.

[3] The verb in Basque agrees with the indirect object as well. In preliminary experiments, the recall of models trained on indirect object prediction was very low, due to the small number of indirect objects in the training corpus; we therefore do not include this task.

verbs before adding our suffixes (for example, by replacing *has* with *have*).

Single Task Results The basic results are summarized in Table 1. Recall is calculated as the proportion of the sentences with a direct object for which the model predicted either SINGULAR or PLURAL, but not NONE. Since all verbs included in the experiment had a subject, subject recall was 100% and is therefore not reported.

Plurality prediction accuracy was higher for subjects than objects. Recall for object prediction was 81.8%, indicating that in many sentences the model was unable to identify the direct object. The lower performance on object plurality prediction is likely to be due to the fact that only about third of the sentences contain a direct object. This hypothesis is supported by the results of a preliminary experiment, in which the model was trained only on transitive sentences (with a direct object). Transitive-only training led to a reversal of the pattern: object plurality was predicted with higher accuracy than subject plurality. We conjecture that this is due to the fact that most noun modifiers in English follow the head, making the head of the object, which in general determines the plurality of the phrase, closer on average to the verb than the head of the subject (see Table 3 below).

The accuracy we report for subject prediction, 94.7%, is lower than the accuracy of over 99% reported by Linzen et al. (2016). This may be due to one of several reasons. First, our training set was smaller: ∼35,000 sentences in our treebank corpus compared to ∼121,000 in their automatically parsed corpus. Second, sentences in the Wall Street Journal corpus may be more syntactically complex on average than sentences in Wikipedia, making it more challenging to identify the verb's arguments. Finally, we predicted agreement in all tenses, whereas Linzen et al. (2016) limited their study to the present tense (where English does in fact show agreement); it may be the case that sentences with past tense verbs are on average more complex than those with present tense verbs, regardless of the corpus.

Multitask Training Accuracy was higher in the joint setting: polypersonal agreement prediction is easier for the model. Subject prediction accuracy rose from 94.7% to 95.7%, object precision was slightly higher (90.0% compared to 88.9%), and object recall was significantly higher, increasing from 81.8% to 85.4%. We hypothesize that supervision signals from the prediction of both arguments lead to more robust abstract syntactic representations that transfer across the two tasks (Enguehard et al., 2017); for example, the model may be better able to identify the head of a noun phrase, regardless of whether it is the subject or the object. These findings suggest that when training on an auxiliary agreement prediction task in order to improve a language model's syntactic performance, additional supervision—in the form of predicting both subject and object—may be beneficial.

4 Order of Core Elements

Languages vary in the typical order of the core elements of a clause: the subject, the object and the verb (Dryer, 2013). For example, whereas in English the canonical order is Subject-Verb-Object (SVO, *The priests are reading the book*), in Irish it is Verb-Subject-Object (VSO, Dillon and Ó Cróinin 1961):

(5)	Léann	[na	sagairt]	[na	leabhair].
	read.PRES	the.PL	priest.PL	the.PL	book.PL

'The priests are reading the books.'

While there are six possible orderings of these three elements, in most human languages the subject precedes both the object and the verb: about 86.5% of the languages use either SOV or SVO orders, 9% of the languages use VOS order, and OVS and OSV languages are extremely rare (Tomlin, 1986).

To test whether RNNs have inductive biases favoring certain word orders over others, we created synthetic versions of English with all six possible orders of core elements. While natural languages often allow at least a limited degree of word order flexibility, our experiments used a simplified setting in which word order was either completely fixed (e.g., always SVO) or fully flexible, where one of the six orders was selected uniformly at random for each sentence in the corpus (the same order is used for all of the clauses of the sentence).

4.1 Corpus Creation

Given a dependency parse for a sentence, we modulated the order of the subject and object nodes with respect to their verb. When changing the position of an argument node, we moved the entire subtree rooted in that node, including verbs and other arguments in this subtree. In the permutation process, we moved to the subject position not only nominal subjects (*nsubj* and *nsubjpass* edges in UD), but also clausal subjects (*csubj* edges). Similarly, we moved to the object position not only nominal objects (*dobj* edge), but also clausal complements (*ccomp* and *xcomp*).

We kept negations, adverbial modifiers, particles and auxiliaries in their original position with respect to the verb. Other non-core dependents of the verb (i.e. not the subject or the object), such as prepositional phrases, were placed according to their original position relative to the verb. For instance, in the clause *the broker took them out for lunch*, the phrase *for lunch* appeared directly following the verb and the arguments of the subtree in which it resides (*took*, *them*, *the broker*) in all word orders, reflecting its original position relative to the verb *took* (see Figure 1). Relative pronouns and complementizers remained in their original position.[4]

In all experiments in this section, we trained the model to jointly predict the plurality of the subject and the object. For consistency across the object and subject plurality prediction tasks, we used the polypersonal agreement markers on all verbs in the sentence (except, of course, for the prediction target, which was withheld completely). For example, in the OVS version of the sentence presented in Figure 1, the input was (6), where *kon* marks the fact that *say* has a plural subject:

(6) them ⟨verb⟩ out frequently the broker for lunch say*kon* they .

4.2 Results

Performance varied significantly across word orders (Table 3). Subject plurality prediction accuracy was inversely correlated with the frequency of attractors (intervening nouns of the opposite plurality) in the language: accuracy was lowest for subject prediction in the VOS and SOV languages, in which objects intervene between the subject and the verb (Figure 2). The degraded performance in these languages is consistent with the attraction effects found in previous studies of agreement in natural languages (Linzen et al., 2016; Gulordava et al., 2018), and support the hypothesis that RNNs have an inductive bias favoring dependencies with recent elements; we test this hypothesis in a more controlled way in §4.3.

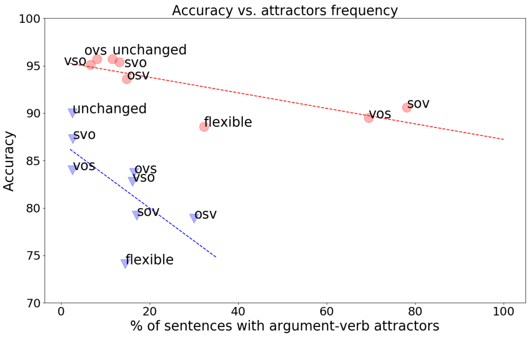

Figure 2: Subject and object plurality prediction accuracy as a function of the percentage of sentences with attractors that are arguments of the verb. Red circles represent subject prediction and blue triangles represent object prediction. R^2: 0.61 for subject, 0.43 for object.

Attractors affected object prediction accuracy as well. The highest accuracy among the synthetic languages was in the SVO language and the worst performance observed in the OSV language. As in §3, subjects were easier to predict than objects, likely because all verbs in the training set had a subject, but only 35% had an object.

Flexible word order was especially challenging for the model, with a subject plurality prediction accuracy of 88.6%, object plurality prediction accuracy of 74.1%, and object recall of 60.2%. This does not necessarily bear on the RNNs' inductive biases: flexible word order without case marking would make it difficult for any learner to infer syntactic relations. Without overt cues, the model must resort to selectional restrictions (e.g., in *the apples ate the man*, the only plausible subject is *the man*), but those are difficult to learn from a small corpus. What's more, some sentences are truly ambiguous when there are no case markers or word order cues; this happens for example when both arguments are animate, as in *the lawyer saw the doctor* (Gibson et al., 2013; Ettinger et al., 2018).

4.3 Withholding Direct Objects in Training

The previous experiments suggested that the RNN has a tendency to identify the more recent argument as the subject, leading to attraction effects

[4] For example, the result of transforming (i) to VSO word order was (ii) rather than (iii):

(i) But these are not the differences that make headlines.

(ii) But are these not the differences that make headlines.

(iii) But are these not the differences make headlines that.

	Subject		Object		
Order	% Attractors	Accuracy	% Attractors	Accuracy	Recall
Unchanged	11.56	95.7 ± 0.23	2.55	90.0 ± 0.1	85.4 ± 2.37
SVO	13.16	95.4 ± 0.41	2.6	87.3 ± 0.23	80.0 ± 2.61
SOV	78.12	90.6 ± 0.37	17.04	79.2 ± 0.78	63.3 ± 4.62
VOS	69.50	89.5 ± 0.54	2.57	84.0 ± 0.39	77.8 ± 3.68
VSO	6.65	95.1 ± 0.12	16.09	82.8 ± 0.7	70.0 ± 1.91
OSV	14.81	93.6 ± 0.23	30.00	78.9 ± 0.17	63.5 ± 4.59
OVS	8.13	95.7 ± 0.37	16.42	83.7 ± 0.32	72.8 ± 1.58
Flexible	32.24	88.6 ± 0.43	14.44	74.1 ± 0.70	60.2 ± 3.24

Table 3: Subject and object plurality prediction for different word orders (recall for the subject is 100% and is not indicated). The % attractors columns indicate the percentage of sentences containing verb-argument attractors. The number are averaged over four runs and the error interval represents the standard deviation.

caused by the object. We conjectured that this is due to the fact that many verbs are intransitive, that is, have a subject but not an object. The clauses in which those verbs appear provide ambiguous evidence: they are equally compatible with a generalization in which the subject is the *most recent* core element before the verb, and with a generalization in which the subject is the *first* core constituent of the clause. Attraction effects suggest that the inductive bias of the RNN leads it to adopt the incorrect recency-based generalization. To test this hypothesis in a controlled way, we adopt the "poverty of the stimulus" paradigm (Wilson, 2006; Culbertson and Adger, 2014; McCoy et al., 2018): we withhold all evidence that disambiguates these two hypotheses (namely, all transitive sentences), and test how the RNN generalizes to the withheld sentence type.

We used the SOV and VOS corpora described before; in both of these languages, the object intervened between the subject and the verb, potentially causing agreement attraction. Crucially, we train only on sentences *without* a direct object, and test on the following three types of sentences:

1. Sentences with an object of the opposite plurality from the subject (object attractor).

2. Sentences with an object of the same plurality as the subject (non-attractor object).[5]

3. Sentences without an object, but with one or more nouns of the opposite plurality intervening between the subject and the verb (non-object attractor); e.g., *The gap between winners and losers will grow* is intransitive, but the plural words *winners* and *losers*, which are a part of a modifier of the subject, may serve as attractors for the singular subject *gap*.

The results are shown in Table 4. Withholding direct objects during training dramatically degraded the performance of the model on sentences with an object attractor: the accuracy decreased from 90.6% for the model trained on the full SOV corpus (Table 3) to 60.0% for the model trained only on intransitive sentences from the same corpus. There was an analogous drop in performance in the case of VOS (89.5% compared to 48.3%). By contrast, attractors that were not core arguments, or objects that were not attractors, did not hurt performance in a comparable way. This suggests that in our poverty of the stimulus experiments RNNs were able to distinguish between core and non-core elements, but struggled on instances in which where the object directly preceded the verb (the instances that were withheld in training). This constitutes strong evidence for the RNN's recency bias: our models extracted the generalization that subjects directly precede the verb, even though the data were equally compatible with the generalization that the subject is the first core argument in the clause.

These findings align with the results of Khandelwal et al. (2018), who demonstrated that RNN language models are more sensitive to perturbations in recent input words compared with perturbations to more distant parts of the input. While in their case the model's recency preference can

[5] When the object is a noun-noun compound, it is considered a non-attractor if its head is not of the opposite plurality of the subject, regardless of the plurality of other elements. This can only make the task harder compared with the alternative of considering compound objects such as "screen displays" as attractors for plural subjects.

	Object (attractor)	Object (non attractor)	Non-object attractor
SOV	60.3 ± 3.7	92.8 ± 0.3	79.2 ± 3
VOS	48.3 ± 2.3	94.0 ± 2.3	83.1 ± 1.1

Table 4: Subject prediction accuracy in the "poverty of the stimulus" paradigm of Section 4.3, where transitive sentences were withheld during training. Numbers are averaged over four runs and the error interval represents the standard deviation.

be a learned property (since recent information is more relevant for the task of next-word prediction), our experiment focuses on the inherent inductive biases of the model, as the cues that are necessary for differentiating between the two generalizations were absent in training.

4.4 Discussion

Our reordering manipulation was limited to core element (subjects, objects and verbs). Languages also differ in word order inside other types of phrases, including noun phrases (e.g., does an adjective precede or follow the noun?), adpositional phrases (does the language use prepositions or postpositions?), and so on. Greenberg (1963) pointed out correlations between head-modifier orders across phrase categories; while a significant number of exceptions exist, these correlations have motivated proposals for a language-wide setting of a Head Directionality Parameter (Stowell, 1981; Baker, 2001). In future work, we would like to explore whether consistent reordering across categories improves the model's performance.

In practice, even languages with a relatively rigid word order almost never enforce this order in every clause. The order of elements in English, for example, is predominately SVO, but constructions in which the verb precedes the subject do exist, e.g., *Outside were three police officers*. Other languages are considerably more flexible than English (Dryer, 2013). Given that word order flexibility makes the task more difficult, our setting is arguably simpler than the task the model would face when learning a natural language.

The fact that the agreement dependency between the subject and the verb was more challenging to establish in the SOV order compared to the SVO order is consistent with the hypothesis that SVO languages make it easier to distinguish the subject from the object (Gibson et al., 2013); indeed, to compensate for this issue, SOV languages more frequently employ case marking (Matthew Dryer, quoted in Gibson et al. 2013).

There was not a clear relationship between the prevalence of a particular word order in the languages of the world and the difficulty that our models experienced with that order. The model performed best on the OVS word order, which is present in a very small number of languages ($\sim$1%). SOV languages were more difficult for our RNNs to learn than SVO languages, even though SOV languages are somewhat *more* common (Dryer, 2013). These results weakly support functional explanations of these typological tendencies; such explanations appeal to communicative efficiency considerations rather than learning biases (Maurits et al., 2010). Of course, since the inductive biases of humans and RNNs are likely to be different in many respects, our results do not rule out the possibility that the distribution of word orders is driven by a human learning bias after all.

5 Overt Morphological Case Systems

The vast majority of noun phrases in English are not overtly marked for grammatical function (case), with the exception of pronouns; e.g., the first-person singular pronoun is *I* when it is a subject and *me* when it is an object. Other languages mark case on most nouns. Consider, for example, the following example from Russian:[6]

(7) a. ya kupil knig-u.
I bought book-OBJECT
'I bought the book.'

b. knig-a ischezla.
book-SUBJECT disappeared
'The book disappeared.'

Overt case marking reduces ambiguity and facilitates parsing languages with flexible word order. To investigate the influence of case on agreement prediction—and on the ability to infer sentence structure—we experimented with different case systems. In all settings, we used "fused" suffixes, which encode both plurality and grammatical function. We considered three case systems (see Figure 1):

1. An unambiguous case system, with a unique

[6] The standard grammatical term for these cases are nominative (for subject) and accusative (for object); we use SUBJECT and OBJECT for clarity.

Case system	Flexible word order		VOS		OVS	
	Subject A	Object A/R	Subject A	Object A/R	Subject A	Object A/R
Unambiguous	99.2 ± 0.5	98.7 ± 0.2 $/98.0 \pm 0.5$	98.9 ± 0.2	99.5 ± 0.1 $/99.1 \pm 0.1$	99.5 ± 0.2	98.6 ± 0.3 $/98.4 \pm 0.6$
Syncretic	99.3 ± 0.2	93.6 ± 0.4 $/88.9 \pm 1.7$	99.1 ± 0.2	97.1 ± 0.2 $/95.0 \pm 1.1$	99.4 ± 0.1	97.8 ± 0.2 $/97.4 \pm 1.2$
Argument marking	96.0 ± 0.3	86.1 ± 0.9 $/79.7 \pm 4.9$	96.9 ± 0.1	93.6 ± 0.1 $/89.8 \pm 2.4$	99.6 ± 0.1	96.8 ± 0.1 $/95.5 \pm 0.5$

Table 5: Accuracy (A) and recall (R) in predicting subject and object agreement with different case systems.

suffix for each combination of number and grammatical function.

2. A partially syncretic (ambiguous) case system, in which the same suffix was attached to both singular subjects and plural objects (modeled after Basque).

3. A fully syncretic case system (argument marking only): the suffix indicated only the plurality of the argument, regardless of its grammatical function (cf. subject/object syncretism in Russian neuter nouns).

In the typological survey reported in Baerman and Brown (2013), 62% of the languages had no or minimal case marking, 20% had syncretic case systems, and 18% had case systems with no syncretism.

Corpus Creation The suffixes we used are listed in Table 2. We only attached the suffix to the head of the relevant argument; adjectives and other modifiers did not carry case suffixes. The same suffix was used to mark plurality/case on noun and the agreement features on the verb; e.g., if the verb *eat* had a singular subject and plural object, it appeared as eat*karker* (the singular subject suffix was *kar* and the plural object suffix was *ker*). We stripped off plurality and case markers from the original English noun phrases before adding these suffixes.

Setup We evaluated the interaction between different case marking schemes and three word orders: flexible word order and the two orders on which the model achieved the best (OVS) and worst (VOS) subject prediction accuracy. We train one model for each combination of case system and word order. We jointly predicted the plurality of subject and the object.

Results and Analysis The results are summarized in Table 5. Unambiguous case marking dramatically improved subject and object plurality prediction compared with the previous experiments; accuracy was above 98% for all three word orders. Partial syncretism hurt performance somewhat relative to the unambiguous setting (except with flexible word order), especially for object prediction. The fully syncretic case system, which marked only the plurality of the head of each argument, further decreased performance. At the same time, even this limited marking scheme was helpful: accuracy in the most challenging setting, flexible word order (subject: 96.0%; object: 86.1%), was not very different from the results on unmodified English (95.7% and 90.0%). This contrasts with the poor results on the flexible setting without cases (subject: 88.6%; object: 60.2%). On the rigid orders, a fully syncretic system still significantly improved agreement prediction. The moderate effect of case syncretism on performance suggests that most of the benefits of case marking stems from the overt marking of the heads of all arguments.

Overall, these results are consistent with the observation that languages with explicit case marking tend to allow a more flexible word orders compared with languages such as English that make use of word order to express grammatical function of words.

6 Related Work

Our approach of constructing synthetic languages by parametrically modifying parsed corpora for natural languages is closely inspired by Wang and Eisner (2016) (see also Wang and Eisner 2017). While they trained a model to mimic the POS tags order-statistics of the target language, we manually modified the parsed corpora; this allows us to control for selected parameters, at the expense of reducing generality.

Simpler synthetic languages (not based on natural corpora) have been used in a number of recent studies to examine the inductive biases of different neural architectures (Bowman et al., 2015; Lake and Baroni, 2018; McCoy et al., 2018). In another recent study, Cotterell et al. (2018) measured the ability of RNN and n-gram models to perform character-level language modeling in a sample of languages, using a parallel corpus; the main typological property of interest in that study was morphological complexity. Finally, a large number of studies, some mentioned in the introduction, have used syntactic prediction tasks to examine the generalizations acquired by neural models (see also Bernardy and Lappin 2017; Futrell et al. 2018; Lau et al. 2017; Conneau et al. 2018; Ettinger et al. 2018; Jumelet and Hupkes 2018).

7 Conclusions

We have proposed a methodology for generating parametric variations of existing languages and evaluating the performance of RNNs in syntactic feature prediction in the resulting languages. We used this methodology to study the grammatical inductive biases of RNNs, assessed whether certain grammatical phenomena are more challenging for RNNs to learn than others, and began to compare these patterns with the linguistic typology literature.

In our experiments, multitask training on polypersonal agreement prediction improved performance, suggesting that the models acquired syntactic representations that generalize across argument types (subjects and objects). Performance varied significantly across word orders. This variation was not correlated with the frequency of the word orders in the languages of the world. Instead, it was inversely correlated with the frequency of attractors, demonstrating a recency bias. Further supporting this bias, in a poverty-of -the-stimulus paradigm, where the data were equally consistent with two generalizations—first, the generalization that the subject is the first argument in the clause, and second, the generalization that the subject is the most recent argument preceding the verb—RNNs adopted the recency-based generalization. Finally, we found that overt case marking on the heads of arguments dramatically improved plurality prediction performance, even when the case system was highly syncretic.

Agreement feature prediction in some of our synthetic languages is likely to be difficult not only for RNNs but for many other classes of learners, including humans. For example, agreement in a language with very flexible word order and without case marking is impossible to predict in many cases (see §4.2), and indeed such languages are very rare. In future work, a human experiment based on the agreement prediction task can help determine whether the difficulty of our languages is consistent across humans and RNNs.

Acknowledgements

This work is supported by the Israeli Science Foundation (grant number 1555/15) and by Theo Hoffenberg, the founder & CEO of Reverso.

References

Matthew Baerman and Dunstan Brown. 2013. Case syncretism. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Mark C. Baker. 2001. *The atoms of language: The mind's hidden rules of grammar*. Basic Books, New York.

Jean-Philippe Bernardy and Shalom Lappin. 2017. Using deep neural networks to learn syntactic agreement. *LiLT (Linguistic Issues in Language Technology)*, 15.

Samuel R. Bowman, Christopher D. Manning, and Christopher Potts. 2015. Tree-structured composition in neural networks without tree-structured architectures. In *Proceedings of the NIPS Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Foris, Dordrecht.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144. Association for Computational Linguistics.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136. Association for Computational Linguistics.

Ryan Cotterell, Sebastian J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* , pages 536–541. Association for Computational Linguistics.

Jennifer Culbertson and David Adger. 2014. Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, page 201320525.

Myles Dillon and Donncha Ó Cróinin. 1961. *Teach Yourself Irish*. The English Universities Press Ltd., London.

Matthew S. Dryer. 2013. Order of subject, object and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Émile Enguehard, Yoav Goldberg, and Tal Linzen. 2017. Exploring the syntactic abilities of RNNs with multi-task learning. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 3–14.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801. Association for Computational Linguistics.

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*.

Edward Gibson, Steven T. Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim, and Rebecca Saxe. 2013. A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, 24(7):1079–1088.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248. Association for Computational Linguistics.

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of language*, pages 73–113. MIT Press, Cambridge, MA.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1195–1205.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of lstms to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231. Association for Computational Linguistics.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 284–294.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint 1412.6980*.

Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. 2018. LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436. Association for Computational Linguistics.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 2879–2888.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1247.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Luke Maurits, Danielle J. Navarro, and Amy Perfors. 2010. Why are some word orders more common than others? A uniform information density account. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.*, pages 1585–1593. Curran Associates, Inc.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2018. Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 2093—2098.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

Shauli Ravfogel, Francis Tyers, and Yoav Goldberg. 2018. Can LSTM learn to capture agreement? the case of Basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107. Association for Computational Linguistics.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

Timothy Angus Stowell. 1981. *Origins of phrase structure*. Ph.D. thesis, Massachusetts Institute of Technology.

Rudolf S. Tomlin. 1986. *Basic word order: functional principles*. Croom Helm, London.

Marten van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 2600–2605, Austin, TX. Cognitive Science Society.

Dingquan Wang and Jason Eisner. 2016. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505.

Dingquan Wang and Jason Eisner. 2017. Fine-grained prediction of syntactic typology: Discovering latent structure with supervised learning. *Transactions of the Association for Computational Linguistics*, 5:147–161.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221. Association for Computational Linguistics.

Colin Wilson. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, 30:945–982.