Do RNNs learn human-like abstract word order preferences?

Richard Futrell[1] and **Roger P. Levy**[2]

[1]Department of Language Science, UC Irvine, rfutrell@uci.edu
[2]Department of Brain and Cognitive Sciences, MIT, rplevy@mit.edu

Abstract

RNN language models have achieved state-of-the-art results on various tasks, but what exactly they are representing about syntax is as yet unclear. Here we investigate whether RNN language models learn humanlike word order preferences in syntactic alternations. We collect language model surprisal scores for controlled sentence stimuli exhibiting major syntactic alternations in English: heavy NP shift, particle shift, the dative alternation, and the genitive alternation. We show that RNN language models reproduce human preferences in these alternations based on NP length, animacy, and definiteness. We collect human acceptability ratings for our stimuli, in the first acceptability judgment experiment directly manipulating the predictors of syntactic alternations. We show that the RNNs' performance is similar to the human acceptability ratings and is not matched by an n-gram baseline model. Our results show that RNNs learn the abstract features of weight, animacy, and definiteness which underlie soft constraints on syntactic alternations.

The best-performing models for many natural language processing tasks in recent years have been recurrent neural networks (RNNs) (Elman, 1990; Sutskever et al., 2014; Goldberg, 2017), but the black-box nature of these models makes it hard to know exactly what generalizations they have learned about their linguistic input: Have they learned generalizations stated over hierarchical structures, or only dependencies among relatively local groups of words (Linzen et al., 2016; Gulordava et al., 2018; Futrell et al., 2018)? Do they represent structures analogous to syntactic dependency trees (Williams et al., 2018), and can they represent complex relationships such as filler–gap dependencies (Chowdhury and Zamparelli, 2018; Wilcox et al., 2018)? In order to make progress with RNNs, it is crucial to determine what RNNs actually learn given currently standard practices; then we can design network architectures, objective functions, and training practices to build on strengths and alleviate weaknesses (Linzen, 2018).

In this work, we investigate whether RNNs trained on a language modeling objective learn certain syntactic preferences exhibited by humans, especially those involving word order. We draw on a rich literature from quantitative linguistics that has investigated these preferences in corpora and experiments (e.g., McDonald et al., 1993; Stallings et al., 1998; Bresnan et al., 2007; Rosenbach, 2008).

Word order preferences are a key aspect of human linguistic knowledge. In many cases, they can be captured using local co-occurrence statistics: for example, the preference for subject–verb–object word order in English can often be captured directly in short word strings, as in the dramatic preference for *I ate apples* over *I apples ate*. However, some word order preferences are more abstract and can only be stated in terms of higher-order linguistic units and abstract features. For example, humans exhibit a general preference for word orders in which words linked in syntactic dependencies are close to each other: such sentences are produced more frequently and comprehended more easily (Hawkins, 1994; Futrell et al., 2015; Temperley and Gildea, 2018).

We are interested in whether RNNs learn abstract word order preferences as a way of probing their syntactic knowledge. If RNNs exhibit these preferences for appropriately controlled stimuli, then on some level they have learned the abstractions required to state them.

Knowing whether RNNs show human-like word order preferences also bears on their suitability as language generation systems. White and Rajkumar (2012) have shown that language gener-

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 50-59.
New York City, New York, January 3-6, 2019

ation systems produce better output when human-like word order preferecences are built in; it may turn out that RNN language models reproduce such preferences such that they do not need to be built in explicitly.

As part of this work, we validate and quantify these word order preferences for humans by collecting acceptability ratings for English sentences with different word orders. To our knowledge, this is the first experimental acceptability-judgment study of these word order preferences using fully controlled stimuli; previous experimental work has used naturalistic stimuli derived from corpora, in which the predictors of word order are not directly manipulated (Rosenbach, 2003; Bresnan, 2007).

Alternations studied

We study four syntactic alternations in English: **particle shift**, in which a verbal particle can appear directly after the verb or later (e.g., *give up the habit* vs. *give the habit up*); **heavy NP shift**, in which a verb is followed by an NP and a PP with order NP–PP or PP–NP; the **dative alternation** (e.g. *give a book to Tom* vs. *give Tom a book*); and the **genitive alternation** (e.g. *the movie's title* vs. *the title of the movie*).

In all these alternations, three common factors influencing word order preferences are evident: short constituents go before long constituents; words which are definite go earlier; and words referring to animate entities go earlier. In fact these preferences are very general patterns across languages, and in some languages constitute hard constraints (Bresnan et al., 2001).

1 Methods

We investigate the learned word order preferences of RNNs by studying the total probability they assign to sentences with various word order properties. Specifically, we create sentences by hand which can appear in a number of configurations, and study how these manipulations affect the SURPRISAL value assigned by an RNN to a sentence. Surprisal is the negative log probability:

$$\begin{aligned} S(x_{i=1}^{n}) &= -\log_2 p(x_{i=1}^{n}) \\ &= -\sum_{i=1}^{n} \log_2 p(x_i | x_{j=1}^{i-1}), \end{aligned}$$

where $x_{i=1}^{n}$ is a sequence of n words forming a sentence and the conditional probability $p(x_i|x_{j=1}^{i-1})$ is calculated as the RNN's normalized softmax activation for x_i given its hidden state after consuming $x_{j=1}^{i-1}$.

Surprisal has a number of interpretations that make it convenient as a dependent variable for examining language model behavior. First, surprisal is equivalent to the contribution of a sentence to a language model's cross-entropy loss: effectively, our RNN language models are trained with the sole objective of minimizing the average surprisal of training sentences, so surprisal is directly related to the model's performance. Second, word-by-word surprisal has been found to be an effective predictor of human comprehension difficulty (Hale, 2001; Levy, 2008; Smith and Levy, 2013); interpreting surprisal as metric of "difficulty" allows us to analyze RNN behavior analogously to human processing behavior (van Schijndel and Linzen, 2018; Futrell et al., 2018). Third, surprisal more generally reflects the dispreference or **markedness** of a sequence according to a language model. High surprisal for a sentence corresponds to a relative dispreference for that sentence. When the logarithm is taken to base 2, surprisal is equivalent to the bits of information required to encode a sentence under a model.

In the studies below, we test hypotheses statistically using maximal linear mixed-effects models (Baayen et al., 2008; Barr et al., 2013) fit to predict surprisals given experimental conditions.

1.1 Models tested

We study the behavior of two LSTMs trained on a language modeling objective over English text: the one presented in Jozefowicz et al. (2016) as "BIG LSTM+CNN Inputs", which we call "JRNN", which was trained on the One Billion Word Benchmark (Chelba et al., 2013) with two hidden layers of 8196 units and CNN character embeddings as input; and the one presented in Gulordava et al. (2018), which we call "GRNN", with two hidden layers of 650 units, trained on 90 million tokens of English Wikipedia.

As a control, we also study surprisals assigned by an n-gram model trained on the One Billion Word Benchmark (a 5-gram model with modified Kneser-Ney interpolation, fit by KenLM with default parameters) (Heafield et al., 2013). The n-gram surprisals tell us to what extent the patterns under study can be learned purely from co-occurrence statistics with a small context window

without any generalization over words. To the extent that LSTMs yield more humanlike performance than the *n*-gram model, this indicates one of two things. Either they have learned generalizations that are formulated in terms of more abstract linguistic features, or they have learned generalizations that can span larger distances than the *n*-gram window.

1.2 Human acceptability ratings

We also compare RNN surprisals against human preferences on our experimental items. We collected acceptability judgments on a scale of 1 (least acceptable) to 5 (most acceptable) over Amazon Mechanical Turk.[1] For the studies of heavy NP shift, the dative alternation, and the genitive alternation, we collected data from 64 participants, filtering out participants who were not native English speakers or who could not correctly answer 80% of simple comprehension questions about the experimental items. After filtering, we had data from 55 participants. For the study of particle shift, we used data from a previous (unpublished) acceptability rating experiment with 196 subjects, and the same filtering criteria. After filtering, we had data from 156 participants.

2 Heavy NP Shift

HEAVY NP SHIFT describes a scenario where constituent weight preferences become so strong that an order which would otherwise be unacceptable becomes more acceptable, as shown in Example (1).

(1) a. The publisher announced a book on Thursday.
b. *The publisher announced on Thursday a book.
c. The publisher announced a new book from a famous author who always produced bestsellers on Thursday.
d. The publisher announced on Thursday a new book from a famous author who always produced bestsellers.

In these examples, the verb *announced* is followed by a noun phrase (*a (new) book...*) and a temporal PP adjunct (*on Thursday*). The usual order for these elements is to put the NP before the PP, but when the NP becomes very heavy, the PP might be placed closer to the verb, which case the word order is called SHIFTED. Heavy NP shift is the primary example of locality effects in word order preferences, in that it creates shorter dependencies from the verb to the NP and the PP.

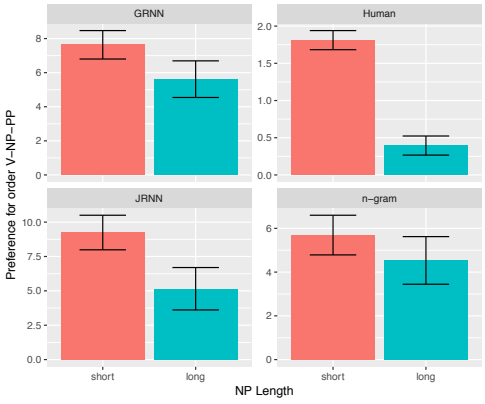

Figure 1: Mean preference for standard word order by NP length. In this and other figures, for computational models, preference is measured as total sentence surprisal for Verb–NP–PP order minus total sentence surprisal for Verb–PP–NP order; error bars represent 95% confidence intervals of the contrasts between conditions, computed by subtracting out the by-item means before calculating the intervals (Masson and Loftus, 2003). For the human data, preference is measured as the difference in mean acceptability for Verb–PP–NP minus Verb–NP–PP, and error bars represent 95% confidence intervals of the contrasts between conditions after subtracting out by-item and by-subject means.

We tested whether RNNs show length-based preferences similar to Example (1).[2] We adapted 40 items from Stallings et al. (1998) which consist of a verb followed by an NP and a temporal PP adjunct, where the order of the NP and the PP and the length of the NP are manipulated. If the networks show human-like word ordering preferences, there should be a penalty for PP–NP order when the NP is short, but this penalty should be smaller or nonexistent when the NP is long.

Figure 1 shows the models' preference for the standard word order (Verb–NP–PP) over the shifted word order (Verb–PP–NP), calculated as the surprisal of sentences in shifted word order minus their surprisal in standard word order. Also included are the human acceptability ratings, where the preference for the order Verb–NP–PP is calculated as the average acceptability difference between Verb–NP–PP and Verb–PP–NP

[1] Preregistered at https://aspredicted.org/sh9zf.pdf.

[2] The preregistration for this experiment can be viewed at https://aspredicted.org/ea6m8.pdf.

across items. In all cases, we see that the shifted order becomes more preferred when the NP is long, although it never becomes the most preferred order.

Our experimental design allows us to control for the effect of sentence length on RNN surprisals. The sentences with long NPs are naturally expected to have higher RNN surprisal than the ones with short noun phrases, since they have more words and thus higher information content. However, the Verb–NP–PP preference is quantified as the *difference* between surprisal of order Verb–PP–NP and surprisal of the order Verb–NP–PP, for both the long and short NP conditions. The long NP conditions may have higher surprisal overall, but this will be cancelled out in the difference. The crucial question is then whether this preference is larger in the long NP case than in the short NP case—whether the red and blue values in Figure 1 are significantly different across items. The crucial statistic for each item i is given by the interaction I_i:

$$I_i = \big(S_i(\text{short, Verb–NP–PP}) - S_i(\text{short, Verb–PP–NP})\big) - \big(S_i(\text{long, Verb–NP–PP}) - S_i(\text{long, Verb–PP–NP})\big),$$

where S_i is the surprisal for the ith item in the given condition. If I_i is significantly positive across items, then we have evidence that NP length causes a preference for Verb–PP–NP order even when controlling for the intrinsic effects of length and of the particular words in each item. The same logic is applied to the analysis of the acceptability ratings data. All studies in this paper apply this same design and analysis. Similar designs are common in psycholinguistics, and are applied to RNN surprisal data in Futrell et al. (2018) and Wilcox et al. (2018).

To test the significance of the interaction, we use mixed-effects modeling with random intercepts and slopes by item. We find that the interaction is statistically significant in JRNN (interaction size 4.0 bits, $p < 0.001$) and GRNN (2.0 bits, $p = 0.02$), but not in the n-gram baseline (1.2 bits, $p = 0.13$). The interaction in JRNN is significantly stronger than in the n-gram baseline ($p = 0.03$), but the interaction in GRNN is not significantly stronger than the n-gram baseline ($p = 0.48$). None of the models show the effect as strongly as the human acceptability judgments.

Thus we find that both JRNN and GRNN exhibit human-like word order biases for Heavy NP shift, but do not find evidence for such a bias in the n-gram baseline. The result suggests that the LSTM models have learned a higher-order generalization that is not trivially present in n-gram statistics.

3 Phrasal verbs and particle shift

Another domain of word order variation similar to Heavy NP shift is phrasal verbs, which consist of a verb and a particle, such as *give up*. The object NP of a transitive phrasal verb can appear in two positions: it can be SHIFTED (after the particle) or UNSHIFTED (before the particle). As in Heavy NP shift, the shifted order is generally preferred when the NP is long:

(2)
a. Kim gave up the habit. [shifted]
b. Kim gave the habit up. [unshifted]
c. Kim gave up the habit that was preventing success in the workplace. [shifted]
d. Kim gave the habit that was preventing success in the workplace up. [unshifted]

The fact that both word orders are possible is called PARTICLE SHIFT. Particle shift provides another arena to test whether RNNs have learned the basic short-before-long constituent ordering preference in English. Furthermore, particle shift is also affected by the animacy of the object NP, in that the unshifted order is preferred when the object NP is animate (Gries, 2003), so we can use this construction to test order preferences involving both length and animacy.

We designed 32 experimental items consisting of sentences with phrasal verbs as in Example (2), where each item could occur with either a long or a short object NP. Long NPs were created by adding adjectives and postmodifiers to short NPs. Half of the items had inanimate objects; half had animate objects. All NPs were definite. We tested the effects of NP length, NP animacy, and word order on language model surprisal.[3]

Figure 2 shows the average preference for shifted word order according to each model, calculated as the surprisal of the shifted order minus the surprisal of the unshifted order. In general, we see that when the object NP is long, the shifted order is relatively preferred; the effect is

[3]The preregistration for this experiment can be viewed at https://aspredicted.org/uu7am.pdf.

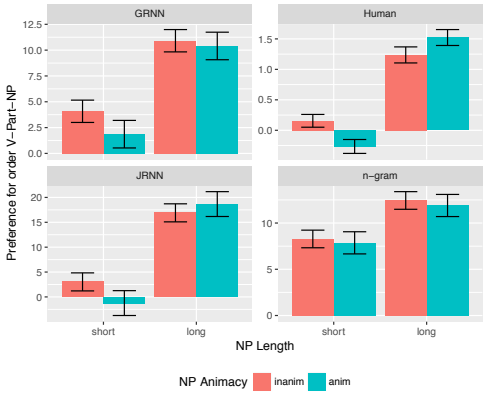

Figure 2: Preference for shifted word order (total sentence surprisal for Verb–Particle–NP order minus Verb–NP–Particle order) by NP length, NP animacy, and model.

strongest in JRNN. In regressions, we found that the interaction of NP length and word order is significant in JRNN (16.9 bits, $p < 0.001$), GRNN (7.8 bits, $p < 0.001$), and the n-gram baseline (4.1 bits, $p < 0.001$). However, the interaction in the n-gram baseline is significantly smaller than in JRNN ($p < 0.001$) and GRNN ($p < 0.01$).

The effects of animacy are unexpectedly intricate. Numerically, GRNN and the n-gram baseline show the expected effect: an animate NP favors unshifted order. However, in the human ratings we find that the expected animacy effect for short NPs actually reverses for long NPs, and this is reflected numerically in JRNN. No effects of animacy are significant in any model. In the human data, animacy has a significant interaction favoring the unshifted word order for short NPs ($p < 0.001$) with an interaction that reverses the effect for long NPs ($p < 0.001$). This reversal is surprising given what was previously known about word order in phrasal verbs.

Our investigation of particle shift has shown that LSTM models learn short-before-long length preferences in regard to word order in phrasal verb constructions; n-gram models show these preferences as well, though weaker. We do not find evidence that the models learned human word order preferences based on NP animacy in this case, but the experimental results suggest that the effects of animacy on this alternation might be more complex than previously believed.

4 Dative alternation

The DATIVE ALTERNATION refers to the fact that in many cases the following forms are substitutable:

(3) a. The man gave the woman the book. [Double-object (DO) construction]
b. The man gave the book to the woman. [Prepositional-object (PO) construction]

The dative alternation is one of the most studied topics in syntax. The two constructions have been argued to convey subtly different meanings, with the DO construction indicating caused possession and the PO construction indicating caused motion (Green, 1974; Oehrle, 1976; Gropen et al., 1989; Levin, 1993). However, the semantic preference of each construction appears to be only one factor among many when it comes to determining which form will be used in any given instance. Other factors include the animacy, definiteness, and length of the THEME (*the book* in Example (3)) and the RECIPIENT (*the woman*) (Bresnan et al., 2007).

The human preferences in the dative alternation work out such that the NP which is more animate, definite, and short goes earlier. An extreme case is exemplified in (4): the sentences marked with ? are relatively dispreferred by native English speakers.

(4) a. The man gave the woman a very old book that was about historical topics.
b. ?The man gave a very old book that was about historical topics to the woman.
c. The man gave the book to a woman who was waiting patiently in the hallway.
d. ?The man gave a woman who was waiting patiently in the hallway a book.

In order to examine whether LSTMs show human-like preferences in the dative alternation, we designed 16 items on the pattern of (3), with 8 verbs of caused possession (such as *give*) and 8 verbs of caused motion (such as *throw*).[4] In all items, the theme was inanimate and the recipient was animate. We manipulated the definiteness of the

[4]The preregistration for this experiment can be viewed at https://aspredicted.org/ky9ne.pdf.

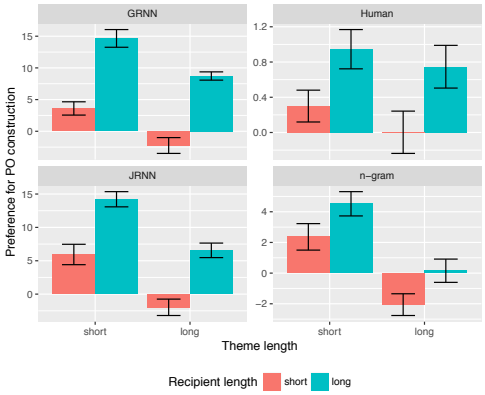

Figure 3: Average PO preference by length of theme and recipient.

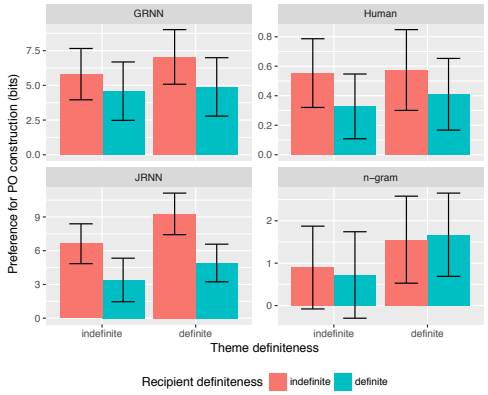

Figure 4: Average PO preference by definiteness of theme and recipient.

theme and the recipient using the articles *the* and *a*, and the length of the theme and the recipient by adding relative clauses to either or both.

Figure 3 shows the strength of the models' preference for the PO construction as a function of the length of the theme and recipient. The LSTMs have an overall preference for the PO form, which is mirrored in the human data. In addition, we see that a long recipient is strongly associated with a stronger preference for the PO form, and a long theme is strongly associated with a relative preference for the DO form, in line with human preferences. The *n*-gram baseline shows these effects but with a smaller magnitude, and without the overall PO preference shown by humans.

All interactions of length and word order are significant at $p < .001$ in all models, with the exception of the effect of recipient length in the *n*-gram model, where $p = 0.01$. The effects of recipient and theme length are significantly weaker in the *n*-gram baseline than in JRNN ($p < 0.01$); for GRNN, the effect of recipient definiteness is significantly stronger than the *n*-gram baseline ($p = 0.02$) but the effect of theme definiteness is not significantly stronger than in the *n*-gram baseline. In human data, the interaction of recipient length and word order is significant at $p < .001$ and the interaction of theme length and word order is significant at $p = .01$.

Now we turn to word order preferences based on NP definiteness. Figure 4 shows the PO preference by the definiteness of the theme and recipient. In line with the linguistic literature, the PO preference is numerically smaller for definite recipients in both LSTM models and in human data, but not in the *n*-gram model. The interaction of recipient definiteness and word order is significant in the expected direction in JRNN at $p < 0.001$ and GRNN at $p = 0.01$. Theme definiteness has a small positive interaction with word order (at $p < 0.01$) for JRNN, favoring the PO construction. These results are broadly in line with the linguistic literature, but they are not reflected in the human data for these experimental items: in the human ratings data, there are no significant interactions of definiteness and word order.

Overall, we find evidence for humanlike ordering preferences in the dative alternation with respect to length and definiteness of theme and recipient. The strongest effects which are most in line with the linguistic literature come from JRNN.

5 Genitive alternation

Similarly to the dative alternation, the GENITIVE ALTERNATION involves two constructions with opposite word orders expressing similar meanings:

(5) a. The woman's house [*s*-genitive, definite possessor]
b. The house of the woman [*of*-genitive, definite possessor]
c. A woman's house [*s*-genitive, indefinite possessor]
d. The house of a woman [*of*-genitive, indefinite possessor]

As in the dative alternation, whatever semantic difference exists between the two constructions is only one factor conditioning which form is used

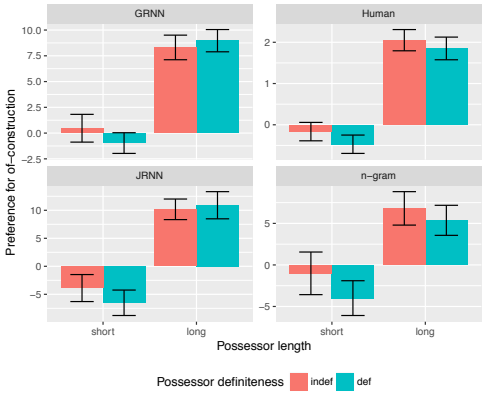

Figure 5: Average *of*-genitive preference by length and definiteness of possessor.

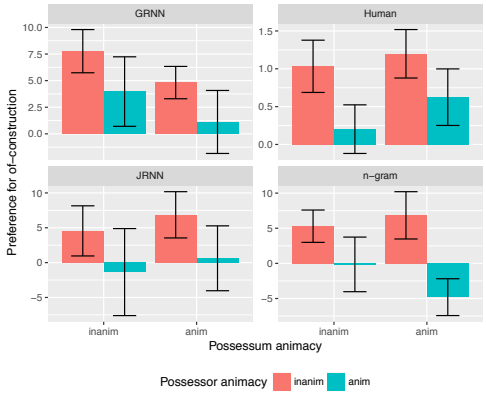

Figure 6: Average *of*-genitive preference by animacy of possessor and possessum.

in each particular case. The other factors are the usual suspects: animacy, definiteness, and length of the POSSESSOR (*the woman* in (5)) and POSSESSUM (*the house* in (5)) (Kreyer, 2003; Rosenbach, 2003, 2008; Shih et al., 2015).

In order to study the genitive alternation in RNNs, we designed 16 items on the pattern of (5). We varied the definiteness and length of the possessor as in the dative alternation.[5] We also varied the animacy of the possessor and possessum, between items.[6]

Figure 5 shows the RNNs' preferences for the *of*-genitive form based on the definiteness and length of the possessor. In all models and in human data, we see that the *of*-genitive is preferred generally when the possessor is long, and the *s*-genitive when it is short. The interaction of possessor length with word order is significant in models and human data ($p < 0.001$ in all cases). Turning to possessor definiteness, we see that it relatively favors the *s*-genitive in human data and in the *n*-gram baseline, in line with the linguistic literature, but no such effect is found in the RNN models. However, the interaction of definiteness with word order is not significant in our data ($p = .09$ in human data and higher for the language models).

Now we turn to effects of animacy. Figure 6 shows the *of*-genitive preference by the animacy of the possessor and possessum. In all models, in line with human preferences, we see that possessor animacy favors the *s*-genitive. The interaction of possessor animacy and word order is significant in JRNN ($p = 0.03$) and GRNN ($p < 0.001$) but not in the *n*-gram baseline ($p = 0.09$); the effect is significantly stronger in JRNN than in the *n*-gram baseline ($p = 0.02$) but not significantly stronger in GRNN. The effect of possessum animacy is more complex: it seems to favor the *s*-genitive in GRNN, but the *of*-genitive in the other models and in human preferences; in any case, the effect is small, and the interaction is not significant in any of the data collected here.

Overall, it appears that the LSTMs tested show humanlike order preferences in the genitive alternation when it comes to possessor length and animacy; they show more evidence for effects of possessor animacy than an *n*-gram baseline. They do not appear to pick up on definiteness preferences, but based on human experimental data these preferences might be weak in the first place.

6 Discussion

We have explored RNN language models' ability to represent soft word order preferences whose formulation requires abstract features such as animacy, definiteness, and length. We found that RNN language models generally do so, and outperform an *n*-gram baseline, indicating that they learn generalizations which are not trivially present in local co-occurrence statistics of words. The extent to which RNNs learn such preferences varies: effects of length are strongly and consistently represented, with weaker evidence for effects of animacy and definiteness.

Much recent work has focused on whether RNNs can learn to represent discrete syntactic

[5] For both constructions to be legitimate syntactic options, the possesssum must be definite and unmodified by relative clauses.

[6] The preregistration for this experiment can be viewed at https://aspredicted.org/f2sk8.pdf.

structures such as long-distance number agreement (Linzen et al., 2016), wh-dependencies (McCoy et al., 2018; Chowdhury and Zamparelli, 2018; Wilcox et al., 2018), anaphora, negative polarity item licensing, and garden path sentences (van Schijndel and Linzen, 2018; Marvin and Linzen, 2018; Futrell et al., 2018). The current work focuses on soft preferences which have been studied in quantitative syntax, and the abstract features that have been discovered to underly these preferences, finding that RNNs are able to represent many of the required features. The same features underlying these soft preferences in English often play a role in hard constraints in other languages (Bresnan et al., 2001): thus our findings indicate that RNNs can learn crosslinguistically useful abstractions.

Our results also demonstrate that some of the key features underlying syntactic alternations can be learned from text data alone, without any particular innate bias toward such features. Qualifying this point, note that the language models we studied here were exposed to many more tokens of linguistic input than a typical child learner.

In addition to the results about RNNs, our work provides human data from directly controlled experimental manipulations of animacy, definiteness, and length in particle shift, genitive, and dative alternations. Our human acceptability ratings experiments have revealed some unexpected patterns, such as the sign reversal in the effect of animacy for long NPs in particle shift (Section 3), which should be investigated in more detail in future work.

An interesting question which has been studied in the functional linguistic literature is *why* these particular word order preferences exist across languages. The preferences are often explained in terms of cognitive pressures on language comprehension and production. The short-before-long preference is most likely a manifestation of the pressure for short dependencies (Wasow, 2002), which is motivated by working memory limitations in sentence processing (Gibson, 1998); this word order preference is reversed in predominantly head-final languages such as Japanese, where there is a general preference for long constituents to come before short ones (Yamashita and Chang, 2001). The biases for animate and definite nouns to come early are usually linked to biases in the human language production process whereby words and constituents which are easier to produce come earlier (Bock, 1982).

It is possible that these same cognitively motivated biases might also be present in RNNs. For example, Futrell and Levy (2017) have argued that there should be a preference for short dependencies in any system that predicts words incrementally given lossy representations of the preceding context: since RNNs represent context using fixed-length vectors, their context representations must be lossy in this way. Furthermore, Chang (2009) has shown that the preference to place animate words earlier can arise in simple recurrent networks without this bias being present in training data, suggesting that RNNs may be subject to similar pressures to produce certain kinds of words earlier.

More generally, we have treated RNN language models essentially as human subjects delivering acceptability judgments, observing their behavior on carefully controlled linguistic stimuli rather than examining their internals. By using controlled experimental designs, we are able to control for factors such as sentence length and the particular lexical items in each sentence (cf. Lau et al., 2017). We believe this approach will allow us to derive initial insight into the limits of what RNNs can do, and will guide work that explains the behavior we document here in terms of network internals.

Acknowledgments

This work was supported in part by a gift from the NVIDIA corporation. RPL gratefully acknowledges support from the MIT-IBM AI Research Laboratory. All code and data is available at https://github.com/langprocgroup/rnn_soft_constraints.

References

R. Harald Baayen, D.J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

J. Kathryn Bock. 1982. Toward a cognitive psychology of syntax: Information processing contributions to

sentence formulation. *Psychological Review*, 89:1–47.

Joan Bresnan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In *Roots: Linguistics in Search of its Evidential Base*, pages 77–96. Mouton de Gruyter, Berlin.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In *Cognitive Foundations of Interpration*, pages 69–94. Royal Netherlands Academy of Science, Amsterdam.

Joan Bresnan, Shipra Dingare, and Christopher D. Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In *Proceedings of the LFG 01 Conference*, pages 13–32. CSLI Publications.

Franklin Chang. 2009. Learning to order words: A connectionist model of Heavy NP Shift and accessibility effects in Japanese and English. *Journal of Memory and Language*, 61:374–397.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

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*, Sante Fe, NM.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Richard Futrell and Roger Levy. 2017. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 688–698, Valencia, Spain.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

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

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

Georgia M. Green. 1974. *Semantics and syntactic regularity*. Indiana Univ Pr.

Stefan Thomas Gries. 2003. *Multifactorial analysis in corpus linguistics: A study of particle placement*. A&C Black.

Jess Gropen, Steven Pinker, Michelle Hollander, Richard Goldberg, and Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in English. *Language*, pages 203–257.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL*.

John T. Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics and Language Technologies*, pages 1–8.

John A. Hawkins. 1994. *A performance theory of order and constituency*. Cambridge University Press, Cambridge.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv*, 1602.02410.

Rolf Kreyer. 2003. Genitive and of-construction in modern written English: Processability and human involvement. *International Journal of Corpus Linguistics*, 8(2):169–207.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Tal Linzen. 2018. What can linguistics and deep learning contribute to each other? *Language*.

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.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of EMNLP*, Brussels.

Michael E. J. Masson and Geoffrey R. Loftus. 2003. Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 57(3):203.

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. *arXiv preprint arXiv:1802.09091*.

Janet L. McDonald, J. Kathryn Bock, and Michael H. Kelly. 1993. Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology*, 25:188–230.

Richard Thomas Oehrle. 1976. *The grammatical status of the English dative alternation.* Ph.D. thesis, Massachusetts Institute of Technology.

Anette Rosenbach. 2003. Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive in english. *Determinants of grammatical variation in English*, 379412.

Anette Rosenbach. 2008. Animacy and grammatical variation: Findings from English genitive variation. *Lingua*, 118(2):151–171.

Marten van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*.

Stephanie Shih, Jason Grafmiller, Richard Futrell, and Joan Bresnan. 2015. Rythm's role in the genitive construction choice in spoken english. In Ralf Vogel and Reuben van de Vijver, editors, *Rhythm in phonetics, grammar, and cognition*, pages 208–234. De Gruyter Mouton, Berlin, Germany.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Lynne M. Stallings, Maryellen C. MacDonald, and Padraig G. O'Seaghdha. 1998. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language*, 39(3):392–417.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

David Temperley and Dan Gildea. 2018. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4:1–15.

Thomas Wasow. 2002. *Postverbal Behavior*. CSLI Publications, Stanford, CA.

Michael White and Rajakrishnan Rajkumar. 2012. Minimal dependency length in realization ranking. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 244–255. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of BlackboxNLP*, Brussels.

Adina Williams, Andrew Drozdov, and Samuel R Bowman. 2018. Do latent tree learning models identify meaningful structure in sentences? *Transactions of the Association for Computational Linguistics*, 6:253–267.

Hiroko Yamashita and Franklin Chang. 2001. "Long before short" preference in the production of a head-final language. *Cognition*, 81(2):B45–B55.