RNN Simulations of Grammaticality Judgments on Long-distance Dependencies

Shammur Absar Chowdhury and Roberto Zamparelli
CIMeC: Center for Mind/Brain Sciences
University of Trento
{shammur.chowdhury, roberto.zamparelli}@unitn.it

Abstract

The paper explores the ability of LSTM networks trained on a language modeling task to detect linguistic structures which are ungrammatical due to extraction violations (extra arguments and subject-relative clause island violations), and considers its implications for the debate on language innatism. The results show that the current RNN model can correctly classify (un)grammatical sentences, in certain conditions, but it is sensitive to linguistic processing factors and probably ultimately unable to induce a more abstract notion of grammaticality, at least in the domain we tested.

Title and Abstract in Italian

RNN Simulazioni di giudizi di grammaticità sulle dipendenze a distanza

L'articolo studia la capacità delle reti neurali LSTM addestrate su un compito di modellazione linguistica di rilevare strutture linguistiche che sono agrammaticali a causa di violazioni nella estrazione di argomenti (dovute alla presenza di argomenti di troppo, o alla presenza di isole del soggetto e delle frasi relative), esplorando le implicazioni per il dibattito sull'innatismo linguistico. I risultati mostrano che l'attuale modello RNN può classificare correttamente frasi grammaticali, in certe condizioni, ma è eccessivamente sensibile a fattori di elaborazione linguistica e probabilmente non in grado di indurre una nozione più astratta di grammaticalità, almeno nel dominio da noi testato.

1 Introduction

Native speaker intuitions about the meaning and grammaticality of linguistic expressions have been the key methodology in theoretical linguistics since at least Chomsky (1957), and are widely seen as a crucial window on the internalized linguistic competence which theoretical linguistics aims to study. Despite lively discussions on the limits of the methodology (see Cowart, 1997; Sprouse and Almeida, 2012; Sprouse *et al.*, 2013, 2016), the availability of very large corpora has not replaced in linguistics the need for judgments on artificially constructed cases whenever theoretical points hinge on the status of very rare or complex constructions, concern languages or dialects for which large corpora do not exist, or involve semantic intuitions that have no easily detectable correlates in corpora (e.g. semantic scope alternations).

Due to its scarce practical applications, modeling grammaticality judgments with computers has never been a typical NLP task, but it can be an important testbed for theories of language processing and grammatical competence. In particular, judgments on syntactic well-formedness require a sensitivity to long-distance structural cues which is a crucial aspect of language competence (Everaert *et al.*, 2015). Elman's (1991) pioneering work on the application of simple recurrent neural network (RNN, Elman 1990) to linguistic sequences showed that such networks could indeed learn some linguistic structures, but had a tendency to forget important linguistic features (e.g. the presence of a Wh-element) as new

Proceedings of the 27th International Conference on Computational Linguistics, pages 133–144
Santa Fe, New Mexico, USA, August 20-26, 2018.

material arrived. However, more recent network architectures, like Long-Short Term Memory Network (LSTMs, Hochreiter and Schmidhuber 1997) or Gated Recurrent Networks (GRU, Chung *et al.* 2014) incorporate memory-management systems which yield much better performances. The seminal work of Linzen *et al.* (2016) and other recent papers (Bernardy and Lappin, 2017; Gulordava *et al.*, 2018) have shown these networks to be capable, in certain conditions, to approximate human levels in a subject-verb number agreement task, even across intervening nouns and verb (e.g. *The boy**s** [that the girl has seen] **are/*is**...*).

These advances raise the question whether similar models, trained on corpora of naturally occurring sentences, could come to approximate the full range of human grammaticality judgments, including judgments on structures which, unlike agreement, are virtually non-existent in the training input. The ability to do so would have implications for the debate on language innatism (Christiansen and Kirby, 2003; Sprouse *et al.*, 2013): a generalist (R)NN has no language-specific learning skills, no innate linguistic parameters to set (Chomsky and Lasnik, 1993; Baker, 2001); if such a device could manage to replicate fine-grained human intuitions inducing them from the raw training input this would be evidence that exposure to language structures (albeit in an amount orders of magnitude larger than the one a child receives, and without a connection to the non-linguistic context of utterance) should in principle be sufficient to derive a syntactic competence, against the innatist hypothesis. Suppose on the other hand that NNs could approximate human intuitions on some linguistic phenomena but not on others, despite similar statistical distributions in the training input: this would now count as strong evidence that the 'unlearnable' phenomena tap on aspects of the grammar faculty that have limited representations in normal language samples, and are good candidates for being innate.

This paper is a step in the direction of this research program. We trained two types of RNNs (GRU and LSTM) on a large corpus of English (Wikipedia and parts of UKWAC, Ferraresi *et al.* 2008), then tested their performances or a range of artificially constructed language structures, contrasting grammatical and ungrammatical examples. Unlike Lau *et al.* (2017), who use LSTM to prove that they can learn the *graded* nature of human grammaticality, we only look at long distance dependencies (the relation between a dislocated element like a Wh-nominal or the head of a relative clause and its gap). As a preliminary task (Task A), we check whether the network is sensitive to the difference in processing complexity between subject and object relatives, a much-studied domain in the psycholinguistic literature; next, we turn to two cases of ungrammaticality, one due to a violation of the principle of Full Interpretation (Chomsky, 1986a, p.98-99) (Task B), one to extraction out of "strong" syntactic islands (Ross, 1982), specifically, subject and relative clause islands (Task C). Violations of these types have been regarded as sharp both in the theoretical (Szabolcsi and den Dikken, 1999) and the experimental literature (Sorace and Keller, 2005; Cowart, 1997; Sprouse *et al.*, 2013).

Our preliminary conclusions are that while the models are able to catch a surprising range of subtle differences in Task A and to correctly classify grammatical and ungrammatical cases in Task B, their performances are highly sensitive to processing factors (esp. sentence length, sentence complexity), a fact which becomes particularly evident in Task C. Two possible conclusions can be drawn: either the NN has troubles disentangling processing and grammaticality, or the most obvious ways for detecting this distinction are not effective for this target.

In the following sections, we first present a brief overview of the methodology used in this study (Section 2), followed by the a detailed task description in Section 3. The results and observation of the network behavior is presented in Section 4; we conclude the study and discuss future directions in Section 5.

2 Methodology

The use of RNN for the evaluation of grammaticality can already be found in Tomida and Utsumi (2013) (a reply to a non-NN learning model in Pearl and Sprouse 2013). However, their RNN (a Jordan RNN, Jordan 1997) worked on abstract data (it was trained on preassigned constituent labels and had to generate other label sequences), a choice that in our opinion requires too many underlying assumptions. RNN trained on raw textual input are first found in Linzen *et al.* (2016), who achieve a very high score (<1%

error rate) in the subject-verb number agreement task. However, they specifically train an LSMN classifier on this task alone. This is fine as a demonstration that the input contains enough network-accessible information to make the correct choice, but it puts the network at an unfair advantage over humans, since giving explicit grammaticality judgment is a rather marginal and non-natural linguistic task (though we make use of *implicit* judgments when we decide if someone is a native speaker, we choose the best wording in a text, etc.). Moreover, it is a task which arguably plays no role in language acquisition. While there are of course immense quantitative and qualitative differences between the language learning process in humans and NNs, we believe that a comparison between their final states can still make sense, under two conditions: (i) that the comparison is not directly between humans and NN judgments, but rather between task-dependent judgment differences, i.e. minimal task-pairs which are very similar for the machine but very different for humans, or vice versa; (ii) that the NN has **not** been specifically trained on this tasks. When Linzen et al. (2016) trained their LSTM on a general task (language modeling: predicting the following word, arguably a more natural task, see van Berkum 2010) and tried to use the resulting network for the judgment task, the error rate increased to around 7% error rate. More recently, a group including Linzen himself (Gulordava *et al.*, 2018) has shown that, with better tuning, a different LSTM trained on a language modeling task is in fact capable of performances comparable to those of the classifier in Linzen *et al.* (2016), even in the absence of any semantic information (a potential confound in Linzen *et al.* 2016).

Assessing grammaticality judgments in a task which is not binary like number agreement raises important methodological questions. We want the network to discriminate between (1a) and (1b). The acceptable Wh-question in (1a) might end with a question mark right at the gap (the object position of *catch*), but also continue with an adverbial, or even (at the coast of decreasing acceptability) a nominal containing a gap (*the tail of*). The ungrammatical case (1b) ends with a normal verb argument (*the tail*).

(1) a. Which mouse did the cat catch {? / last night ? / the tail of ? }
b. *Which mouse did the cat catch the tail?

The question is which NN measure best corresponds to the speaker's perception of ungrammaticality in (1b), keeping into account that even in the theoretical and psycholinguistic literature there are no established metrics to measure 'degrees of ungrammaticality' (see Cowart 1997; Sorace and Keller 2005 for discussion).

2.1 Evaluation Measures

We explored various possibilities, from *global* measures like *perplexity* (PPL) and non-normalized sentence cross-entropy loss (CEL), to *local* measures like the normalized log probability of a full stop $LogP_n(FS)$ or question mark $LogP_n(QM)$ after the current word.

Cross-entropy loss (CEL) measures the distance between the output distribution predicted by the model and one-hot vector representing the target word (t_i). The loss for the test sentence can therefore be described as an approximation of the equation $CEL = \sum_{i=1}^{n} \ln P(t_i)$, where $P(t_i)$ is the probability given by the model to the i^{th} target word in the sentence of length n. In our analysis, we mostly use the averaged CEL (ACEL) while comparing each particular set of cases. Note that (A)CEL is not normalized by sentence length. Perplexity measures how many equally probable words can follow a point in the text; as the sentence grows longer and more information accumulates, the options for the following word decrease. Perplexity is calculated by $PPL = e^{-\frac{TCEL}{N}}$, where the total cross-entropy loss ($TCEL = \sum CEL$) is computed for the sub-dataset and the total number of words, N, in the corresponding dataset. Both measures are based on the intuition that an ungrammatical sentence should 'confuse' the NN more than a corresponding grammatical one, and that this confusion will translate in a decreased ability to make correct predictions.

As for the local measure, normalized log probability of the full stop and question mark is calculated as $LogP_n(QM/FS) = log(\frac{p_m(\Im)}{p_u(\Im)})$ where $p_m(\Im)$ is the probability of the symbol, $\Im$, at a given position given by the model; $p_u(\Im)$ is the unigram probability of the $\Im$.

As it turns out, neither PPL or CEL are perfect ways to evaluate a language model for ungrammatical-

ity, since they do not locate it at a specific point in the sentences. Yet, this could also be an advantage, since global measures like PPL/CEL can potentially catch parsing problems that arise earlier than expected, and record a perturbation in the NN's later predictions as it recovers from an ungrammatical point earlier in the text. Therefore, these methods seem appropriate for a first exploration of this task.

On the other hand, the advantage of the local measure used in this study (P(FS/QM), i.e. the confidence that the sentence is about to end) is that it can give precise information about the point at which ungrammaticality is detected (e.g. after *catch* in (1)), or can be used to track the NN expectations through time, as we do in Figure 1. The disadvantage is that, as example (1a) shows, a sentence can always continue in unexpected ways. Moreover, in some cases of ungrammaticality the presence of individual words might not be a significant predictor.

2.2 RNN Architecture

In order to simulate an "unbiased learner" that tries to model human grammatical intuitions, inducing them from the raw training input, we designed a RNN-based language model. A language model is often defined as a model learning the conditional probability over words, given a historical context. The model includes a vocabulary V of size $|V|$; when learning a training sentence, each word is represented by a one-hot encoding vector, $x \in \mathbb{R}^{|V|}$, with a corresponding index in V.

For this study we used two successful variant of simple RNN – long-short term memory (LSTM) and gated recurrent unit (GRU) models. GRU is basically an LSTM without the output gate; the content of the memory cell is copied into the network at each time steps. We trained the models varying the number of hidden units (u=$\{$100, 1500$\}$). For our datasets we observed that the un-tuned LSTM (u=100) performed slightly better than the GRU architecture with same parameters. Though it is possible that with better tuning the GRU could outperform LSTM, for the purpose of this study we present data from just the LSTM model tuned only for number of hidden units $u = 1500$, $layer = 2$ and embedding dimension, $e = 256$. To avoid over-fitting the model on the training data we also applied dropout — a regularization technique — in different layers. For training the model, we used a PyTorch RNN implementation with an SGD optimizer. We have not tuned the models for different dropout or learning rate parameters, and we used a fixed batch size of 80.

To train the RNN model we used an English corpus extracted from Wikipedia plus a small subset of UKWaC ($\approx$ 0.32% of the training data), crawled from *.uk* domain. For the Wikipedia data, we downloaded the WikiText-103 raw data containing the original case, punctuation and numbers. We then tokenized both datasets, removing urls, email addresses, emoticons and text enclosed in any form of brackets ($\{.\}$,(.), [.]). We replaced rare words (tokens with frequency $\leq$10) with *<unk>* token along with its signatures (e.g. *-ed, -ing, -ly* etc.) to represent every possible out-of-vocabulary (OOV) words. We also replaced numbers (exponential, comma separated etc) with a *<NUM>* tag. After these preprocessing steps, our training data consisted of $\approx 136M$ words, with a vocabulary of size $|V| = 0.1M$. It is to be noted that the style of the texts selected to train the language model is mostly encyclopedic, due to the prevalence of Wikipedia over other web corpus data. This is of course not representative of the typical registers of English language, but it does give us a good proportion of complex embedded sentences, which match some of the deeply embedded constructions present in the task set. Recall, moreover, that the measures in this study are always relative, i.e. they contrast minimally different inputs within the same language model. Thus, any imbalance in the training data should not be expected to make a large difference in the global results.

3 Task Description

In order to filter out the effect of words which could affect the network performance in ways orthogonal to the structures at issue, we opted to increase the number of sentences to evaluate by building them as *sentence schemata*, (e.g. (2) for the subject relative Task A), which were expanded automatically to generate all the possible ways of picking one of the expressions within $\{\}$. Note that some of the variable were experimental conditions (e.g. the presence of *that* or *who* in (2)), others were added just to increase variety at the level of the content lexicon, so as to minimize possible effects of collocations or sentences

the network might have encountered in the training phase. The results we present are averaged across all the sentences that express the same experimental condition.

(2) { The / A / Every / Some } { student / man / professor / driver } { that / who } had { seen / spoken with / interviewed / mentioned / approached / met / lived with }{ her / Mary / the woman } gave a brief speech.

The phenomena[1] we tested are the following:

A. SUBJECT VS. OBJECT RELATIVE CLAUSES: Much psycholinguistic literature since Gibson (1998); Gordon *et al.* (2001) has shown that object relatives such as (3a) are harder to parse than subject relatives (3b) (in terms of time, accuracy, etc.), and that the presence of certain material intervening between the head (*boy*) and the gap (indicate with an underscore) can affect reaction times (Gordon *et al.*, 2002, 2004). While this is not a contrast in grammaticality (both structures are clearly acceptable for adults, Adani *et al.* 2010), we designed this preliminary task to check if the network was sensitive to the position of the gap and to the type of intervening subject: a pronoun, a proper name or a full noun phrase. According to Rizzi's Featural Relativized Minimality (Friedmann *et al.*, 2009; Adani *et al.*, 2010), an intervener with many grammatical features in common with the moved element can make the extraction harder, or even ungrammatical (see esp. Villata *et al.* 2016 for the case of 'weak islands').

(3) a. The boy that Mary has invited __ *object extraction*
b. The boy that __ invited Mary *subject extraction*

A second variable was the relative pronoun, which could be *that*, *who* or null (in object relatives: *The girl Anna saw*). This test set comprised 1680 expanded sentences.

B. WH EXTRACTIONS: A second test set involves cases of Wh-extraction where the gap position could be empty (4a) or filled by an overt element (4b) (a personal pronoun, an indefinite pronoun, a demonstrative NP).

(4) a. Which candidate/issue should the students discuss ?
b. *Which candidate/issue should the students discuss {him / it / something else / this candidate / this issue}

(4b) is a very strong semantic/syntactic violation, since either the Wh or the final nominal cannot be connected to the verb, violating the principle of Full Interpretation (Chomsky, 1986b). However, certain uses of pronouns are standardly treated as bound variables in formal semantics (and there are languages, e.g. Hebrew or Welsh, which use so-called *resumptive* pronouns in place of gaps at least in relative clauses), so we expected that filling the gap with pronouns might be better than filling it with full noun phrases. If the NN is able to carry over semantic information from the Wh-phrase to the gap position, we also expected that gap-fillers that match in animacy (*which candidate . . . him/this candidate*) should be better than non matching cases (*which issue . . . him/this candidate*). The sentences were generated at 0, 1 or 2 levels of embedding (e.g. *Who did John claim [that the professor assumed [that the students should discuss (it)]]?* is level 2), to see if distance makes the network 'forget' the Wh or its features.

All the Wh cases above were contrasted with corresponding affirmative sentences. In this case, however, the gap is the ungrammatical case (5a), while the gap fillers we see in (5b) all yield grammatical sentences. The pre-gap verbs, *discuss, mention, describe, write about, worry about, address, promote, consider* were chosen not to easily allow intransitive counterparts.

(5) a. *The student should consider .

[1]The expanded test sets for each task can be found in
https://github.com/LiCo-TREiL/Computational-Ungrammaticality.

b. The student should consider {him / it / something else / this candidate / this issue}

This allows us to directly compare the interrogative and affirmative case. What makes this task particularly interesting is the fact that, locally, the beginning and the end of each sentence is perfectly grammatical; they become strongly ungrammatical only when seen together, possibly at a distance which is rarely (if ever) attested in corpora. When fully expanded, this test set contains 72720 sentences.

C. SUBJECT AND RELATIVE ISLAND VIOLATIONS: While the previous test measures the ability of the network to carry information from the Wh phrase across the whole sentence, the last test set pitches them against the phenomenon of *syntactic islands* (Ross, 1967). Descriptively, a *subject island* blocks a dependency whose gap is *inside* (hence the tag 'subextraction') a nominal subject (6a), as opposed to a nominal object (6b); a *relative clause island* bars gaps inside relatives (7). Note that (7b) combines the two types of islands, and should be worse if the effect of multiple violations is cumulative.

(6) a. *Who did [a classmate of __] ruin John ?
b. ?Who did John see of [a classmate of __]?

(7) a. *Which girl did John see [the person that dated __]?
b. **Which girl did [the person that dated __] see John ?

Despite decades of research, there is no established functional explanations why islands exist (though see Szabolcsi and Zwarts 1990; Sprouse *et al.* 2012 for some approaches). Thus, they represent a good starting point to verify the limits of NN performance. In this case, we provided Y/N question and affirmative counterparts for each of the Wh-interrogatives, as a point of reference.

4 Results and Discussion

Task A. Subject vs. Object Relative Clauses: Tables 1-2 contains the results of the Subject *vs.* Object Relative Task. As it can be observed, the network is better (both in terms of PPL and CEL) at dealing with subject than object relatives, thus indicating a sensitivity for the position of the gap which corresponds to the preference found in humans, and especially children (Adani *et al.*, 2010). Moreover, the network loss (Avg. CEL) improves (i.e. it is lower) when the other nominal inside the RC is a pronoun (recall that in *object* relatives the non-gapped nominal intervenes between the Wh-element and the gap): the pronoun improves by ≈ 2.83 over (proper name) PN and ≈ 2.94 over the (noun phrase) NP in subject RC, by ≈ 5.55 and ≈ 5.93 respectively in object relative; there is no significant difference in loss between PNs and full NPs. The fact that the pronoun effect is larger in subject position is particularly noteworthy. It could be due to the greater frequency of pronouns in subject than object position reported in Gordon *et al.* (2001), but this tendency is also in line with the Featural Relativized Minimality approach of Friedmann *et al.* (2009); Villata *et al.* (2016): in object RC the RC-internal nominal intervenes between the relative head and its gap; the more features the head and the nominal have in common, the more the connection between head and gap is disrupted; pronouns have fewer features in common with the relative head (*boy* in (3)) than other nominals (e.g. no +N(oun)), so they interfere less.

With respect to relative pronoun type, Table 2 shows a significant preference for *that* over *who* in Subject RC, and a preference for null relative pronouns over overt ones in object RC (i.e. *the boy Mary saw* > *the boy that/who Mary saw*, ">" = easier for the NN), in terms of ACEL. This seems in line with corpus frequencies measured on UKWAC01, where **Det NP** *that* is about 4 times more frequent than **Det NP** *who*.

Cases	**PPL**	**ACEL** ($\pm std$)	**#**
Subj-relatives	84.52	55.99 ($\pm$3.60)	672
Obj-relatives	105.59	57.25 ($\pm$4.23)	1008

Table 1: Task A Results: Subject vs. Object RC. # represents the number of instances in that sub-data

Nom-Rel	Subj-R *ACEL*	Obj-R *ACEL*	Rel-Pronoun	Subj-R *ACEL*	Obj-R *ACEL*
pronoun	54.07	53.42	*that*	57.7	57.4
proper name	56.9	58.97	*who*	54.28	58.25
full Noun Phrase	57.01	59.35	*no Rel Pron.*	–	56.09

Table 2: Task A Results: Nom-Rel represents Nominal inside Relative clause and Rel-Pronoun represents Relative Pronouns.

Task B. Wh extractions and FI violations: To analyze the Task B dataset we initially divided it into four cases: [WH. . . GAP], [AFF(ERMATIVE). . . NOGAP] (both grammatical) and [WH. . . NOGAP], [AFF. . . GAP] (both ungrammatical). We studied the overall network perplexity (PPL) and the ACEL loss for the sentence given by the NN. We observed that the PPL is 106.10 overall for the grammatical sets, 151.57 for the ungrammatical ones. When we keep Wh and affirmative cases apart, [AFF. . . NOGAP] have PPL = 67.72 (calculated over 11640 instances), which is as expected lower than [AFF. . . GAP], PPL = 76.63 (calculated over 2940 instances). However, the perplexity given by grammatical [WH...GAP] (163.16, with 11616 instances) is higher than that of the ungrammatical [WH...NOGAP]-sentences (PPL = 156.42, 46560 instances).

Options to Select from	CEL	Comments
O1: What should Mary discuss? *Gramm.*	**29.14**	
O2: What should Mary discuss it?	32.79	
O3: What should Mary discuss him?	34.32	CEL(O1) <CEL(O2-6)
O4: What should Mary discuss something else?	39.63	Correct if O1 is chosen
O5: What should Mary discuss this topic?	40.22	
O6: What should Mary discuss this candidate?	42.37	
O1: The professor has said that Mary should consider. *UnGramm.*	**48.40**	
O2: The professor has said that Mary should consider it.	49.15	CEL(O1) <CEL(O2-5)
O3: The professor has said that Mary should consider something else.	55.53	Correct if any of O2-O5
O4: The professor has said that Mary should consider this topic.	56.87	is chosen
O5: The professor has said that Mary should consider this candidate.	58.40	

Table 3: Example of Gramm. vs UnGramm. classification task by RNN using ACEL as a measure. Lower values are better.

To explore this further, we designed a simple classification task where each sentence was presented with 5 or 6 possible alternatives in the gap position, as shown in Table 3. The NN's choice was correct if the CEL for the correct option was the lowest. The experiment included a total of 14520 instances containing Wh and affirmative sentences and had an accuracy of 91.45%, indicating that the RNN network is largely able to pick the correct option for Wh (99.1% out of 11616 instances) and for Aff (60.7% out of 2904 instances). The very different margins of the two effects, which probably account for the WH/AFF PPL difference above, are actually not unexpected, as it is probably easier to accommodate the existence of an intransitive version of a transitive verb than to explain away an extra argument.

Since all the sentences in Task B could potentially end at the (filled) gap, we decided to investigate the effect of sentence type on the network's perception that the sentence is about to end. Figure 1 tracks the NN expectation that the following word is going to be a full stop (P(FS)), or a question mark (P(QM)). The results are intriguing: in general, an end-of-sentence (EOS) is least expected after an auxiliary or modal; in affirmatives "." is unlikely after the final verb, very likely after the object. Interestingly, the possibility of a question mark (i.e. a question marked by intonation alone) is always present in affirmatives. Wh-sentences are dominated by the expectation of a gap (whose proxy is "?"), peaking at the first verb and decreasing slowly. The unexpected arrival of the object seems to convince the NN that the sentence is after all not a direct interrogative ("." higher than "?").

Turning to an analysis of the impact of each filled overt element in the Wh-extraction dataset, Table 3 shows that when the gap is filled by a personal pronoun (Pro, e.g. *it, him*), the overall PPL, of such set, is much better than when the gap filled by an indefinite pronoun (IndPro, *something else*) or a demonstrative NP (a 10 point difference). A similar pattern obtains with ACEL: Pro $>$ IndPro $>$ DemNP. The ability of the NN to track semantic information (specifically, animacy) from the Wh to the gap (i.e. *which candidate . . . this candidate* and *which issue . . . this topic*) was also confirmed: the global PPL of the cases with a match in animacy is lower ($PPL = 220.05$) than the unmatched cases ($PPL = 223.89$),

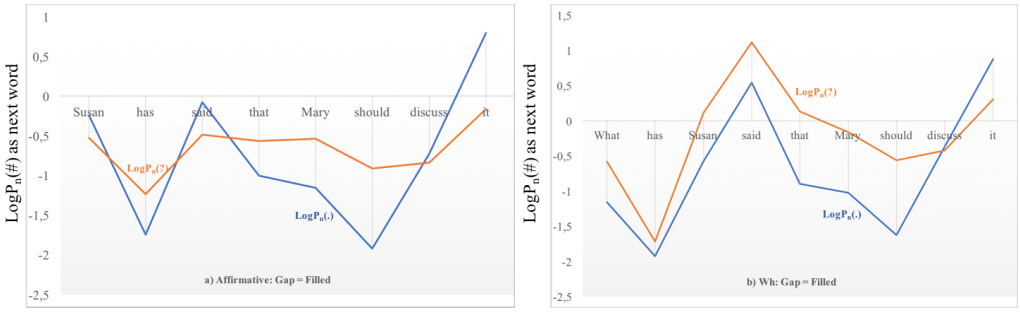

Figure 1: The EOS expectations for different types of sentences. Log Probability that the next word is a "?" (red) or a "." (blue), according to the LSTM model. The sentence is an example for the whole category.

but by only 3 points; once again, a similar pattern with a minor difference of 0.31 obtains using average CEL.

In Figure 2, we presents the effect of different levels of embedding between the wh-element and the gap position in cases where the gap is filled (Ungrammatical cases) or empty (Grammatical). Our findings suggests that in all our scenarios (grammatical and ungrammatical, Wh and affirmatives) an increase in the level of embedding increases the average CEL very significantly. A similar pattern is observed with affirmative sentences.

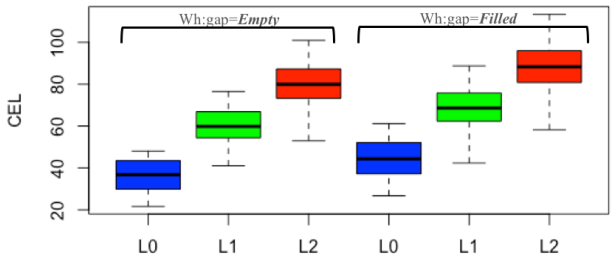

Figure 2: Variation of CEL for different level of embedding; Wh cases with gap empty or filled.

Task C. Subject and Relative Island Violations: So far we have seen that the RNN model was able to distinguish grammatical from ungrammatical pairs with some success, but also to capture a number of interesting effect from the psycholinguistic literature: in Task A, the preference for subject relatives and the effect of intervening pronouns vs. full DPs (see Table 2); in Task B, the almost acceptability of Wh-resumptive pronouns vs. indefinites vs. the fully ungrammatical demonstrative fillers (Table 3), the matching animacy, and the shifting preferences for full stops or question marks (Figure 1). However, in many of these cases the margin of success was small with respective to others; for instance, only 3 ACEL points divide *What should Mary discuss* from *What should Mary discuss it*, but the ACEL distance between e.g. *What should Mary discuss* and *What has she said that Mary should discuss* is over 20 ACEL points.

We now turn to see how the model performs on the typically syntactic Task C. The overall results for subject (S) and object (O) position are presented in Table 4. Since many of the gaps are non final we do not employ the local P(FS/QM) measure and only present the two global measures. At a first glance, the network seems to have been able to capture the facts with exactly the right progression: in the ACEL measure object subextractions are best, followed at an 8 ACEL point distance by subextractions from subjects, subject of passives, relatives in object position and relatives in subject position. The PPL facts are harder to fit, especially for the passive (NSp: 95.64). It should be noted, however, that NSp could be interpreted as a *parasitic gap* construction (Ross, 1967; Engdahl, 1985), which is normally judged fairly acceptable (*Who was a portrait of __ painted by __ ?*), plus a spurious *Tim* filling the second gap, as in Task B.

Has our model learned syntactic islands from Wikipedia? It would appear so, down to the fact that the

Cases		PPL	ACEL ($\pm std$)	#
OBJ NP extraction (NO):	*Who has Tim seen a portrait of __?*	95.87	43.50 ($\pm$8.32)	480
SUBJ NP extraction (NS):	*Who has a portrait of __ scared Tim?*	159.78	48.37 ($\pm$7.30)	480
SUBJ NP extr. from pass. (NSp):	*Who was a portrait of __ painted by Tim?*	95.64	49.75 ($\pm$5.95)	480
OBJ RC extraction (RO):	*Who has Tim seen a portrait that showed __ ?*	131.75	51.64 ($\pm$8.76)	768
SUBJ RC extraction (RS):	*Who has a portrait that showed __ scared Tim?*	181.87	56.11 ($\pm$6.90)	960

Table 4: Overall results of Task C: Subject and Relative Clause (RC) island violations, where overall PPL represents the perplexity on the sub-dataset, ACEL represents the average cross-entropy loss given by the network and # represents the number of instances in that sub-data set. The losses differ from each other significantly, p-value < 0.05.

Cases	Wh-interrogatives	Y/N interrogatives	Affirmatives	Ratio PPL	Ratio CEL
	ACEL	ACEL	ACEL	(Wh-/YN), (Wh-/Aff)	(Who/YN), (Who/Aff)
OBJ pos. extraction (NO)	**51.26**	**45.81**	**36.4**	(1.57, 2.78)	(1.12, 1.42)
SUBJ. pos. extraction (NS)	54.32	51.05	41.73	(1.27, 2.03)	(1.06, 1.30)
SUBJ. Pos. extr. + Passive (NSp)	53.51	51.92	45.08	(1.07, 2.01)	(1.03, 1.19)
OBJ. RC (RO)	59.11	54.35	43.95	(1.43, 2.46)	(1.09, 1.34)
SUBJ. RC (RS)	60.76	58.54	50.57	(1.12, 1.47)	(1.04, 1.20)

Table 5: Results of Task C: Subject and Relative Island Violations, divided in to *wh- interrogatives*; *Y/N interrogatives* and *affirmatives*. ACEL represents the average cross-entropy loss given by the network and #, the number of instances in that sub-data set. For ACEL the lower the value the better. The losses we obtain differ from each other significantly, p-value < 0.05.

last case, RS extraction, is the worst because, arguably, it is the sum of two distinct islands — relative and subject. However, a look at the performance of the model on a set of parallel cases shows that the interpretation of these facts should probably be quite different. Recall that for each Wh case, our Task C test set contains the corresponding Y/N and affirmative sentences (8):

(8) a. Who did John see the person that dated __?
b. Did John see the person that dated Mary?
c. John saw the person that dated Mary.

Table 5 shows the ACEL scores for all three types, as well as the ratio between the scores given to the sentence types. One remarkable aspect is that the Wh cases have a much higher (i.e. worse) score than the corresponding affirmative; even more remarkable, however, is the fact that Y/N questions are also very far from assertions, much closer in fact to their Wh counterparts. But Y/N questions have no Wh gap, hence no island effects. Equally remarkably, Y/N-questions and assertions follow a progression which is almost identical to the one of Wh-cases: NO>NS>NSp>RO>RS. This is quite visible from the ratios, which remain very stable under ACEL, and taper down slightly in PPL.

This data shows that the increased perplexity with Wh cases has nothing to do with island effects, or we would not find it in Yes/No questions and assertions. Our hypothesis is that it is rather the cumulative effect of increasing syntactic complexity, plus position. Suppose that an NP such as *a classmate of John* is more complex than *John*, *a classmate* and possibly *John's classmate*, thus potentially more ambiguous. Suppose further that relative clauses are even more complex/ambiguous. Ambiguity leads to uncertainty, so by increasing it we increase perplexity as well, yielding the difference between NO,NS>RO,RS.

This does not yet explain why NO>NS and RO>RS (recall that they are made of the same words, just in different orders). This, we hypothesize, is the effect of position. A complex structure at the beginning of a sentence (subject position) can be more damaging than one at the end (object position), probably since it can lead the intermediate network units into a "wrong" state of activation, which will be fed back to the RNN as the next word enters, generating additional perplexity. To test this hypothesis, we placed other types of complex nominals (NP conjunctions, NPs containing adjectival and PP modifiers in both subject ((9a), (9c)) and object position (e.g. (9b), (9d)) in Y/N-interrogatives and in the corresponding affirmative sentences. We observed that the PPL/ACEL for sentences with complex subjects was 154.79/73.01, vs. 99.70/66.64 with complex objects. The pattern was similar in Y/N-interrogatives

and affirmatives, and across the two different types of complex nominals.

(9) a. Did [the publisher, the journalist and their families] meet Mary?
b. Did Mary meet [the man, the woman and their families]?
c. Did [a well built, blond-haired man with a large, heavy backpack] meet Mary ?
d. Did Mary meet [a well built, blond-haired man with a large, heavy backpack]?

In addition, all things being equal, a Wh might generate additional perplexity, since it leads the model to expect a gap at multiple points (cf. Figure 1b). Putting these factors together, it is not strange that the RS case (relative complexity + subject position + Wh perplexity) might come out with the highest score.

5 Conclusions and future work

Our work builds on Linzen *et al.* (2016); Gulordava *et al.* (2018); Bernardy and Lappin (2017), but with a difference goal: the possibility for an RNN to learn to classify sentences by grammaticality, focusing on two specific long-distance effects: the presence of an extra argument (Task B) and the island effects in extraction (Task C). The results showed, first of all, that our NN model was good at capturing known effects in the processing of relatives (subject/object preferences, effect of interveners), and good at spotting the selective need for a final argument (Task B). The fact that the ungrammatical cases were graded, with pronouns next best after gaps (see Table 3) shows that the network wasn't just using the simple rule "If it starts with Wh*, pick the version with fewer words, if not, don't". On the other hand, the overwhelming effect of processing factors like the level of embedding (Figure 2), and the fact that the apparent success of the NN in the island task in not based on the island extraction effect itself cast doubts on the idea that the NN is using an abstract dimension of 'grammaticality'.

It could be tempting to take this as a cue that even human ungrammaticality should be reduced to processing (see Villata *et al.* 2016 for discussion in the domain of *weak islands*, which we did not test here), but there are reasons to believe that, while processing might play a role, it cannot be the whole story. As is well-known to anyone who has practiced a musical instrument, pronounced tongue-twisters or read centrally-embedded sentences, processing difficulties improve a lot with practice. However, multiple repetitions of *Who did John see Mary?* by humans are not likely to make it better.

The study raises several methodological issues, first and foremost in the choice of measures. Perplexity seems to be an obvious first choice for a language model, but even when it is normalized by the number of words in the sentence it is sensitive to contrasting effects: longer sentences are more predictable, but recursively embedded sentences are less. Moreover, complex structures increase perplexity more when they are at the beginning of a sentence than at the end (see our discussion of Task C results). In future work we plan to explore variations of these measures with different properties, and use *bidirectional* LSTMs to mitigate the latter effect. Local prediction tracking as in Figure 1, on the other hand, seems to be a promising, intuitive tool to see what the NN is "thinking" throughout a parse, but it is not always applicable. This is related to a final point. Grammaticality is typically judged relative to a (grammatical) point of comparison. Thus, *Who did a portrait that showed scared Tim?* is truly awful if judged as the RC+Subj extraction from *The portrait that showed **who** scared Bill?*, but what if the NN takes it as a variant of a structure such as *[(the person) who has a dress that (really) showed] scared Tim*? Humans can be given minimal pairs to make the difference clear. Learning how to do this with NN remains for future work. Only after these issues have been resolved and a performance plateau has been reached we will be in a position to go back to the original question: are (R)NN feasible models of innate-grammar-free language learners? Which abstract properties can they learn from the input?

References

Adani, F., van der Lely, H. K., Forgiarini, M., and Guasti, M. T. (2010). Grammatical feature dissimilarities make relative clauses easier: A comprehension study with italian children. *Lingua*, **120**(9), 2148 – 2166.

Baker, M. C. (2001). *The atoms of language: the mind's hidden rules of grammar.* Oxford University Press, Oxford.

Bernardy, J.-P. and Lappin, S. (2017). Using deep neural networks to learn syntactic agreement. *LiLT*, **15**(2).

Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hague.

Chomsky, N. (1986a). *Barriers*. Linguistic Inquiry Monograph 13. MIT Press, Cambridge, Mass.

Chomsky, N. (1986b). *Knowledge of Language: Its Nature, Origins and Use*. Praeger, New York.

Chomsky, N. and Lasnik, H. (1993). The theory of principles and parameters. In J. Jacobs and al., editors, *Syntax: An International Handbook of Contemporary Research*, volume 1, pages 506–569. Walter de Gruyter.

Christiansen, M. H. and Kirby, S. (2003). Language evolution: consensus and controversies. *TRENDS in Cognitive Sciences*, **7**(7), 300–307.

Chung, J., Cho, C. G. K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments*. Sage Publications, Thousand Oaks.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, **14**, 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, **7**(2–3), 195–225.

Engdahl, E. (1985). Parasitic gaps, resumptive pronouns, and subject extractions. *Linguistics*, pages 3–44.

Everaert, M., Huybregts, M., Chomsky, N., Berwick, R., and Bolhuis, J. (2015). Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, **19**(12), 729–743.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop WAC4*, pages 47–54.

Friedmann, N., Belletti, A., and Rizzi, L. (2009). Relativized relatives: types of intervention in the acquisition of a-bar dependencies. *Lingua*, **119**(1), 67–88.

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

Gordon, P., Randall, H., and Marcus, J. (2001). Memory interference during language processing. *J. Exp. Psychol. Learn. Mem. Cogn.*, **27**, 1411–1423.

Gordon, P., Hendrick, R., and Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language*, **51**(1), 97–114.

Gordon, P. C., Hendrick, R., and Levine, W. H. (2002). Memory load interference in syntactic processing. *Psychological Science*, **13**, 425–430.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

Jordan, j. M. (1997). Serial order: A parallel distributed processing approach. *Advances in psychology*, **495**, 121–471.

Lau, J. H., Clark, A., and Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, **41**(5), 1202–1241.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 521–535.

Pearl, L. and Sprouse, J. (2013). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, **68**, 20–23.

Ross, J. R. (1967). *Constraints on Variables in Syntax*. Indiana University Linguistics Club, Bloomington.

Ross, J. R. (1982). Pronoun deleting processes in German. Paper presented at the Annual Winter Meeting of the LSA, San Diego.

Sorace, A. and Keller, F. (2005). Gradience in linguistic data. *Lingua*, **115**, 1497–1524.

Sprouse, J. and Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's "core syntax". *Journal of Linguistics*, **48**(3), 609–652.

Sprouse, J., Wagers, M., and Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. *Language*, **88**(1), 82–123. Linguistic Society of America.

Sprouse, J., Schütze, C. T., and Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001-2010. *Lingua*, **134**, 219–248. DOI: 10.1016/j.lingua.2013.07.002.

Sprouse, J., Caponigro, I., Greco, C., and Cecchetto, C. (2016). Experimental syntax and the variation of island effects in english and italian. *Natural Language and Linguistic Theory*, **34**, 307–344.

Szabolcsi, A. and den Dikken, M. (1999). Islands. *GLOT Internationaal*, **4**(6), 3–8.

Szabolcsi, A. and Zwarts, F. (1990). Semantic properties of composed functions and the distribution of wh-phrases. In M. Stokhof and L. Torenvliet, editors, *Proceedings of the Seventh Amsterdam Colloquium*, pages 529–555. Institute for Language, Logic and Information, Amsterdam.

Tomida, Y. and Utsumi, A. (2013). A connectionist model for acquisition of syntactic islands. *Procedia - Social and Behavioral Sciences*, **97**, 90–97.

van Berkum, J. J. A. (2010). The brain is a prediction machine that cares about good and bad – any implications for neuropragmatics? *Italian Journal of Linguistics*, **22**.

Villata, S., Rizzi, L., and Franck, J. (2016). Intervention effects and relativized minimality: New experimental evidence from graded judgments. *Lingua*, **179**, 76–96.