Recurrent babbling:
evaluating the acquisition of grammar from limited input data

Ludovica Pannitto
CIMeC
University of Trento
ludovica.pannitto@unitn.it

Aurélie Herbelot
CIMeC/DISI
University of Trento
aurelie.herbelot@unitn.it

Abstract

Recurrent Neural Networks (RNNs) have been shown to capture various aspects of syntax from raw linguistic input. In most previous experiments, however, learning happens over unrealistic corpora, which do not reflect the type and amount of data a child would be exposed to. This paper remedies this state of affairs by training a Long Short-Term Memory network (LSTM) over a realistically sized subset of child-directed input. The behaviour of the network is analysed over time using a novel methodology which consists in quantifying the level of grammatical abstraction in the model's generated output (its 'babbling'), compared to the language it has been exposed to. We show that the LSTM indeed abstracts new structures as learning proceeds.

1 Do RNNs learn grammar?

Artificial Neural Networks, and Long Short-Term Memory Networks more specifically, have consistently demonstrated great capabilities in the area of language modeling. In addition to generating credible surface patterns, they show excellent performances when tested on very specific grammatical abilities (Gulordava et al., 2018; Lakretz et al., 2019), without requiring any prior bias towards the syntactic structure of natural languages.

From a theoretical point of view, these results seem to contradict the well-known argument of the *poverty of the stimulus* (Chomksy, 1959; Chomsky, 1968) and raise questions about the *continuity hypothesis* in language acquisition (Lust, 1999; Crain and Pietroski, 2001). At the same time, a number of results give a much more mitigated view of RNNs' abstraction capabilities (Marvin and Linzen, 2018; Chowdhury and Zamparelli, 2018). It thus remains unclear how and to what extent grammatical abilities emerge in artificial language models, and how this knowledge is encoded in their representations – especially when considering notions such as *productivity* and *compositionality* (Baroni, 2020), which are recognised as defining traits of natural languages.

This paper proposes that the evaluation of RNN grammars should be widened to include the effect of the type of input data fed to the network, as well as the theoretical paradigm used to analyse its output. We specifically remark that much of the discussion concerning language modeling remains influenced by the mainstream *generativist* approach, which posits a sharp distinction between syntax and the lexicon. Our own approach will be to depart from this account by testing the grammatical abilities of an RNN in a *usage-based* perspective. Specifically, we ask what kind of structures are abstracted and used productively by the network, and how the abstraction process takes place over time.

In contrast with previous models: (i) we train a vanilla char-LSTM on a more realistic variety and amount of data, focusing on a limited amount of child-directed language; (ii) we do not rely on extrinsic evaluations or downstream tasks, instead we introduce a methodology to evaluate how the distribution of grammatical items, over time, comes to approximate the one in the input, through a continuous process and (iii) we tentatively explore the interaction between meaning representations and the abstraction abilities of the network, blurring the distinction between lexicon and syntax, in a way more akin to Construction Grammar (CxG, Fillmore, 1988; Goldberg, 1995; Kay and Fillmore, 1999). Our evaluation focuses on the network's generated output (its 'babbling'), asking to what extent the system simulates the type of grammatical abstraction observed in human children. The study is conducted on English.

In what follows, we review related work (§ 2), we then formulate the question of grammar modelling in a broader theoretical framework (§ 3) in-

volving three parameters: the type of acquisition mechanism under study, the nature of the input data, and the representational paradigm adopted for the analysis. We configure this broad framework with particular choices of parameters and implement it in § 4, 5 and 6. We provide two analyses of the distributional properties of the network's 'babbling', discussed in § 7.

2 Related Work

A considerable amount of literature has investigated the ability of ANNs to acquire grammar, and the list we present here is by no means exhaustive. The analysis of the syntactic abilities of LSTMs (Hochreiter and Schmidhuber, 1997) and ANN-based language models dates back quite a few years (McClelland, 1992; Lewis and Elman, 2001). Recent contributions have followed a general tendency to analyze the inner-workings of networks, and the specific type of knowledge they acquire (Alishahi et al., 2019; Linzen and Baroni, 2020). For instance, Linzen et al. (2016) show how a network acquires *abstract* information about number agreement, albeit in a supervised setting. The same study is expanded in Gulordava et al. (2018), which shows how a language modeling task is enough for a network to predict long-distance number agreement, both on semantically sound and nonsensical sentences. The authors conclude that "LM-trained RNNs can construct abstract grammatical representations", but their model is trained on a rather consequent amount of data (90M tokens) from a rather peculiar distribution (a Wikipedia snapshot). Similarly, it has been shown that LSTMs (McCoy et al., 2018; Wilcox et al., 2018) can learn tricky syntactic rules like the English auxiliary inversion and filler-gap dependencies, although, in later work, McCoy et al. (2020) find that only models with an explicit inductive bias (Shen et al., 2018) learn to generalize the MOVE-MAIN rule with respect to auxiliary inversion. Marvin and Linzen (2018) show instead poor performance of RNNs in grammaticality evaluation, due to their sensitivity to the specific lexical items encountered during training, a limitation that, they say, "would not be expected if its syntactic representations were fully abstract". Similarly Chowdhury and Zamparelli (2018) state that their model "is sensitive to linguistic processing factors and probably ultimately unable to induce a more abstract notion of grammaticality". Moreover, despite the fact that the

model of Gulordava et al. (2018) is tested on four languages, the most promising results may not be generalizable to languages showing different surface patterns from English. Ravfogel et al. (2018) fail to replicate Gulordava et al. (2018)'s results on Basque, and Davis and van Schijndel (2020), after testing the network on relative clause attachment cases in English and Spanish, conjecture that the associative (non-linguistic) bias of RNNs overlaps with English syntactic structure but represents an obstacle to learn attachment rules for Spanish.

Other puzzling results concern the relation of perplexity to syntactic performance (Warstadt et al., 2019; Hu et al., 2020): having evaluated their models on 34 benchmarks, Hu et al. (2020) conclude with a call for a wider variety of syntactic phenomena to test on. Further studies have shown that networks carrying explicit inductive bias perform better than vanilla LSTMs. In a recent paper, Lepori et al. (2020) show that a constituency-based network generalizes more robustly than a dependency-based one, and that both outperform a more basic BiLSTM. Lastly, we mention the study carried out by Kuncoro et al. (2018) who perform their study using a character-based LSTM – a choice we will follow in this work.

A very similar scientific discussion, which we won't report in depth here, is blooming around Transformer-based language models (Tran et al., 2018; Goldberg, 2019; Bacon and Regier, 2019; Jawahar et al., 2019; Lin et al., 2019), leading to similar contrasting results.

Finally, a separate line of work focuses on a more indirect test of the information encoded in the internal representation, assessing which aspects of the original syntactic structure can be reconstructed through diagnostic classifiers (Adi et al., 2017; Giulianelli et al., 2018; Hewitt and Manning, 2019; Tenney et al., 2019).

In summary, a clear trend has not yet emerged (Linzen and Baroni, 2020). All the models we cited, however, seem to idealize syntactic structure as a separate and more abstract ability from the knowledge of statistical regularities or lexical co-occurrences. This perspective may reflect a belief in a sharp distinction between the *lexicon* and *compositional rules*. That is, ANNs are expected to gain *abstract grammatical abilities* through compositional generalization, where compositionality is understood as the ability to produce an unbounded number of sentences by means of a

set of algebraic rules (Baroni, 2020). In contrast with this approach, usage-based models encourage us to adopt a different perspective, and to analyze LSTMs' grammatical abilities with respect to the kind of representations (more in §3.3) posited by theories such as Construction Grammar (CxG, Fillmore, 1988; Goldberg, 1995; Kay and Fillmore, 1999).

3 Framework

In essence, the question of language acquisition asks how much language (Λ) can be learned with a certain level of computational complexity (C) by being exposed to a certain type of data (I). The corresponding formalization, $a : C \times I \mapsto \Lambda$, describes both human and artificial acquisition processes, and its components have been central in the linguistic debate. Below, we will discuss each term (C, I and Λ) in further detail.

3.1 Computational complexity of the acquisition mechanism (C)

Our aim is to test how much grammatical structure can be induced from linguistic input through a pattern-finding mechanism such as that provided by ANNs. Therefore, we fix the level of computational complexity to a vanilla, character-based LSTM, which we train exploring different sources of input in a specific range $\{I_i\}$, selected based on their complexity level. We then use the trained model to generate some amount of text (to *babble*), to explore the structure of the produced output $\ell \in \Lambda$, mainly with respect to productivity.

$$(LSTM, I_i) \xrightarrow{a} \ell_i \qquad (1)$$

Our choice of model has consequences from a theoretical point of view. Different stances have been taken about how much has to be *hard-coded* or *innate* in order for language acquisition to happen: while formal innatist theories have always posited the need for a specialized and innate ability, a dedicated *device* for language learning (Chomsky, 1981, 1995; Hauser et al., 2002), cognitive theories have argued for a more systemic vision, showing how general purpose memory and cognitive mechanisms can account for the emergence of linguistic abilities (Tomasello, 2003; Goldberg, 2006; Christiansen and Chater, 2016; Cornish et al., 2017; Lewkowicz et al., 2018).

LSTMs, under this perspective, can be seen as a domain-general attention and memory mecha-nism, without any explicitly hard-coded grammatical knowledge. They have been applied, without substantial modifications, to a variety of tasks, ranging from time series prediction to object co-segmentation, and encompassing grammar learning as well. On the continuum between specialized devices and general purpose associative mechanisms, LSTMs place themselves on the latter side, with their recurrent structure seeming to be crucial in the linguistic abstraction process (Tran et al., 2018).

3.2 Structure and role of the input (I)

Because of the traditional sharp distinction between *competence* and *performance*, the role of the input and the linguistic environment has been minimized by theories in the realm of Universal Grammar (UG). Usage-based theories, on the other hand, have granted the input a central role to the end of explaining why language is structured as it is (Fillmore, 1988; Kay and Fillmore, 1999; Hoffmann et al., 2013; Christiansen and Chater, 2016; Goldberg, 2019): one of the striking points to make here is that in the usage-based framework the acquisition problem is framed as an incremental process. Acquiring language essentially entails learning how to process the linguistic input in an error-driven procedure, where full linguistic creativity and productivity are acquired gradually by speakers (Bannard et al., 2009), building up on knowledge about specific items and restricted abstractions.

In this sense, the specific features of the language on which ANNs are trained cannot be overlooked when it comes to describing their acquired grammatical abilities. Compared to what a child is exposed to during the most crucial months of language acquisition, ANNs are trained on an input that is often unrealistic in size: the LSTM introduced in Gulordava et al. (2018) is for example exposed to 90M tokens, and sees them multiple times over training. It is hard to come up with a precise estimate of the amount of language children are exposed to during the years of acquisition, as the variation depends on a huge number of factors including the socio-economic environment (Bee et al., 1969) or the societal organization (Cristia et al., 2019). Hart and Risley (1995), in a seminal work, estimate that, by the age of 3, welfare children have heard about 10 millions words while the average working-class child has heard around 30 millions. Finally, the domain of the data also matters: child-directed language is characterized

by specific features ([Matthews and Bannard](), [2010]) that are not present in the most widely used corpora.[1]

3.3 Shape and features of the generated language (Λ)

Any analysis of the language Λ generated by a learner implies the availability of a *representation*. Much has been written on the respective benefits of various representations of linguistic structures: the exact nature of their shape and content is the ultimate conundrum of linguistic theory. Of course, this paper is not the place to review the wide variations that exists among theories, so we will just limit ourselves to motivate our choice with respect to the broader theoretical framework.

Constituency-based representations have been prevalent in the description of natural language syntax, becoming primarily associated with derivational theories. Due to the Fregean view of compositionality, they have also become the natural building blocks for meaning composition. Dependency representations have, on the other hand, re-gained popularity over constituency representations in the last decades, showing desirable properties from a computational perspective (they adapt to a wider array of languages, representing ill-formed sentences results easier and the output is more easily incorporated in semantic graphs) while taking a more functional approach to language description, more in line with cognition oriented-approaches.

In order to represent the features of $\ell \in \Lambda$, we choose a representation which makes the least possible assumptions on the acquisition process and on the content of the generated language, and is at the same time flexible and computationally tractable. We therefore rely on dependency representations, more specifically the universal dependencies framework ([Nivre et al.](), [2020]), from which we extract subtrees called *catenae* ([Osborne et al.](), [2012]). As we will see below, the notion of catena is more flexible than that of constituent, and allows us to describe a larger set of generalizations.

Generally speaking, CxG approaches seem to lack a shared representational framework[2], relying on box diagrams or Attribute-Value Matrices to describe the traits of the fragments they study. The

Strings	A, AB, ABC, ... B, BC, ...E
Catenae	A, B, C, D, E, AB, ABCE, ABDE, ABCDE, ABE, BCE,BDE, BE, CE, DE, CDE
Constituents	A, ABCDE, C, D, CDE

Table 1: Possible structures that can be extracted from the dependency tree in Figure [2]

structures introduced by [Osborne]() ([2006]) are characterized instead as fundamental meaning-bearing units ([Osborne and Groß](), [2012]), in line with the theoretical tenets of CxGs, thus being ideal candidates for the lexicon (or *'Constructicon'*) postulated in such theories: catenae have in fact been applied in the description of construction-like structures ([Osborne and Groß](), [2012]; [Dunn](), [2017]) and allow for the representation of non-adjacent structures while encompassing the notion of constituent as well ([Osborne](), [2006], [2018]).

A catena is defined as "a word, or a combination of words which is continuous with respect to dominance" ([Osborne et al.](), [2012]): given a dependency tree, this definition selects a broader set of elements than the definition of constituent[3]. Unlike constituents, catenae can include both contiguous and non contiguous words. They however capture something more refined than generic subsets of sentence items, as the elements are grouped depending on the syntactic links holding in the sentence.

From a graph-theory perspective, catenae form subtrees (i.e., subsets of nodes and edges that constitute a tree themselves) of the original tree.

Let's consider for example the structures represented in Figures [1a], [1b] and [1c]: the same elements (nodes A to G) are arranged differently in the structure of dependency tree, and this leads to a different number and composition of catenae.

As a concrete example, Figure [2] represents a dependency tree, and Table [1] the structures that can be extracted from it: considering the lexical level, we can extract *Mary had lamb*, *had a lamb*, *a little lamb* as catenae. As the morpho-syntactic and syntactic levels are available, however, we can also extract partially filled structures as *Mary had NOUN*, *nsubj VERB dobj* and so on.

Of interest for our analysis, CxG argues that grammar items above the lexical level bear meaning themselves, and that this emerges from patterns of usage. According to [Goldberg]() ([2006]), for example, the meaning of the ditransitive pattern *Sbj V*

---

[1]Specifically, those that contain data harvested from the web such as Wikipedia or UKWaC.

[2]an exception should be made for the formalisms derived from the FrameNet project (https://framenet.icsi.berkeley.edu/)

[3]which can be seen as a subtype of catena as "A catena that consists of a word plus all the words that that word dominates"

(a) The case of a flat structure, where all nodes are linked to the *root*: from a tree like this we can extract $2^6 - 1$ catenae, each one containing A plus a subset of its children nodes.

(b) The case where nodes are arranged in a full dependency chain: here the number of catenae corresponds to the number of substrings that could be extracted from the linear signal, that is 20.

(c) The case of a hierarchical structure, typically what we would find in linguistic trees, where the counts are less trivial to make. In particular, for each node we find that the number of catenae *rooted* in that node can be estimated depending on the number of catenae *rooted* in his children nodes, and depends therefore on the specific structure of the tree.
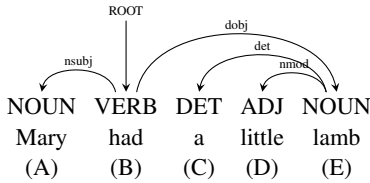
Figure 1

Figure 2: The dependency representation of the sentence *Mary had a little lamb*, annotated with morphosyntactic and syntactic information.

Obj Obj2, and thus its productivity, emerges from its strong association with *give* in child-directed speech: part of the meaning of *give* remains attached to the construction. A natural, and promising (Rambelli et al., 2019), solution to represent the semantics of catenae is given by Distributional Semantics (Harris, 1954), where each element of the *'Constructicon'* is implicitly described in terms of its context of use (Erk, 2012; Lenci, 2018). We will see in §6 how we can use such distributional representations to investigate the level of abstraction of our network's babbling.

4 Data and language modelling

4.1 Corpus

Our corpus is composed of three parts, each presenting different features with respect to linguistic complexity: (1) Child-directed utterances of the publicly available North American and United Kingdom portions of the CHILDES database (MacWhinney, 2000); (2) movie and TV series subtitles from the OpenSubtitle corpus (Lison and Tiedemann, 2016), filtered by content-rating label (G for movies and TV-Y, TV-Y7, TV-G for TV series), available from *The Movie Database*[4]; (3) a 2019 snapshot of Simple English Wikipedia[5], an English-language edition of Wikipedia written in basic English.

These different corpora vary in size: for our experiments we randomly (with uniform probability) extract sentences from each source so that the total number of tokens approximates 3 millions (10% are kept for validation and 10% for testing).

4.2 Language models

For each of the considered corpora, we train a character-based LSTM on the tokenized, raw text. To do so, we slightly modify the PyTorch implementation of a vanilla LSTM.[6] adapting it to a character-based setting. We run a Bayesian optimization process (Nogueira, 2014–) to select the best hyperparameters for the corpus (values can be found in the supplementary material). We then produce a model every 5 epochs of training (for a total of 7 models for CHILDES, 9 models for Open Subtitles and 7 models for simple Wikipedia), as to be able to produce *snapshots* of the network's abilities at different stages during training. For each of the saved models, we sample[7] utterances until we reach the size of the input (the 'babbling' stage). An example of babbling is reported in Table 2.

4.3 Extracting catenae

As introduced in § 3, the outcome of the acquisition process is a language sample ℓ_i, that we want to compare to the input language I_i or to other language samples ℓ_j produced at different stages of acquisition. For the next steps, both the input text

---

[7]The sampling happens as follows: a random initial letter is picked, with a probability depending on the distribution of letters at the beginning of sentences in the input data, then letters are sampled with a greedy algorithm until an *end of sentence* marker is reached or the length surpasses the average sentence length of the input plus 2 standard deviations.

	CHILDES	**opensubtitles**	**simplewiki**
input	you tinker tot let 's see if I can make this turn here that 's Jim 's business	this is no way to treat a lion ! re-entry into earth 's atmosphere in 37 minutes . are you worried i 'm going to try and stop you ?	She is the rector of the National Autonomous University of Honduras (UNAH) since 2009 .
best model	she 's a fire if I put it down for a snack oh I love you	some of horases are here down . just lost it all . i said we 've ... lucky .	She is a former municipality in the center of an arrondissement in the southwest of France .

Table 2: Examples from input text and babbling produced by the best model, for each corpus. Sentences have been sampled according to the distribution of sentence lengths in the data.

(the corpus) and the network's babbling are linguistically processed and annotated up to the syntactic level with the UDPipe toolkit (Straka and Straková, 2017) (a schema of the full processing pipeline is presented in Figure 3). Since our aim is to monitor the syntactic behaviour of the network throughout learning, we extract catenae from the input corpus and from each babbling stage. To do so, we perform a recursive depth-first visit of dependency trees (pseudocode is provided in the supplementary material). That is, if the node A is a leaf, then the only possible catena is the one containing A itself; otherwise, all catenae rooted in A are formed by A plus a (eventually empty) combination of catenae rooted in its children nodes.

With this procedure, we extract catenae from sentences (with length between 1 and 25). For efficiency reasons, we exclude catenae longer than 5 elements. Many structures are generated, not all of which are relevant: since we see catenae as pieces of the lexicon, frequency is not the only relevant parameter and elements should be positively associated in order to be recorded as objects. We therefore weigh the produced structures with a multivariate version of Mutual Information (MI), based on Van de Cruys (2011):

$$MI(x_1, ..., x_n) = f(x_1, ..., x_n) \log_2 \frac{p(x_1, ..., x_n)}{\prod_{i=1}^{n} p(x_i)} \quad (2)$$

where $p(x_1, ..., x_m) = \frac{f(x_1, ..., x_m)}{\sum_{(y_1, ..., y_m)} f(y_1, ..., y_m)}$.

Table 3 shows some of the structures with highest and lowest MI: from a qualitative perspective, it is evident that the measure is able to isolate linguistically relevant patterns, such as the basic intransitive and transitive structures (@*nsubj* @*root* and @*nsubj* _VERB @*obj*).

It is important to remark that the linguistic annotation process (except for the tokenization step) and the catenae extraction processes are completely independent from the language modeling performed

catena	frequency	mi
largest mi		
@nsubj @root	294.59K	633.93K
_DET _NOUN	189.97K	552.32K
_VERB @obj	190.72K	520.82K
_PRON _VERB	271.44K	503.17K
@nsubj _AUX @root	129.60K	478.86K
smallest mi		
_PRON @nsubj	17.50K	-35.54K
@root @nsubj	27.61K	-34.89K
@nsubj _PRON	11.63K	-30.47K
_VERB @nsubj	12.79K	-26.82K
_AUX _PRON	15.75K	-26.67K

Table 3: Examples of catenae extracted from CHILDES. Largest and smallest mutual information are reported, in top and bottom tier of the table respectively. Part of Speech are prefixed by "_" and syntactic relations are prefixed by "@"

by the LSTM, which is only fed with raw text and is therefore completely agnostic about the linguistic categories superimposed by the parser.

5 What do ANNs approximate?

Our first analysis demonstrates that the language generated by the LSTM reproduces the distribution of the input, and that this happens well beyond the lexical level: in other words, the network has acquired statistical regularities at the level of grammatical patterns, and is able to use them productively to generate novel language fragments that adhere to the same distribution as the input.

Fig. 4 shows the extent of this approximation for various pairs: (i) $(\ell_i^c, \ell_j^c) \in \ell_{1...k}^c$ (language fragments output by a particular stage of babbling, for each corpus c), (ii) $(\ell_i^c, I^c), \ell_i^c \in \ell_{1...k}^c$ (fragments output by a particular stage of babbling, compared to those extracted from the respective input c), (iii) $(I^{c_i}, I^{c_j}), (BM^{c_i}, BM^{c_j}), (I^{c_i}, BM^{c_j})$ (fragments extracted from the input or the best babbling stage, compared among different corpora c_i, c_j). It emerges from the plot that correlations are very high within each corpus (on average, 0.935 for
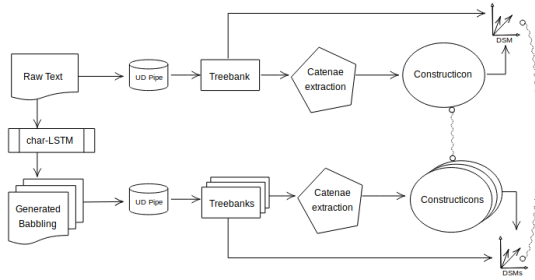
Figure 3: The figure depicts the processing pipeline used for the experiments: raw text from corpora serves as input to the LSTM, that in turn produces raw text at different training stages (i.e., the babbling). Both the corpus and the babbling texts are then processed with a NLP pipeline in order to build treebanks, from which catenae are then extracted. These extracted structures form the constructicons, which are compared in the experiments described in Section 5 (dashed line). The structures in each constructicon are then represented in a Distributional Semantic Model, through co-occurrences extracted from the respective treebanks. The distributional semantic models are then used for the experiments in § 6 (dashed line).

CHILDES, 0.929 for OpenSubtitles and 0.917 for Simple Wikipedia). In particular, the correlations between the best models (BM) and the respective input series (I) show values that are among the highest, demonstrating that the network acquires structures and reproduces them with a distribution that almost perfectly matches the input. On the other hand, it is clear that different corpora show different distributions, as correlations between pairs of input series I and best models show much lower values[8]. Overall, CHILDES scores the best correlation values, probably due to the specific features of child-directed speech, specifically its repetitiousness Clark (2009). Open-Subtitles interestingly shows intermediate properties, sharing quite a lot of catenae with CHILDES,[9] while Simple Wikipedia shows a completely different distribution.

6 Meaning and abstraction

Our second analysis relies on the idea that we can state that the network *has learned some grammar* once it is able to use an acquired pattern in a pro-
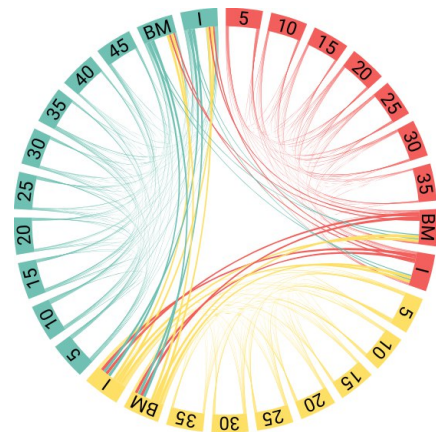
Figure 4: Correlation values (Spearman ρ) over top 10K catenae for each corpus (OpenSubtitles in green on the left of the plot, CHILDES in red in the top right and Simple Wikipedia in yellow at the bottom) compared to the respective babbling (at intermediate stages of learning) and the best models (BM). The thickness of the connections is inversely proportional to correlation.

ductive and creative way. Following the basic hypothesis of CxG, stated in § 3.3, we expect this generalization ability to evolve during training and the distributional properties of patterns to be in relation with the grammatical abilities of the network at various stages of learning.

Let's consider the structures cat_1 : *the dog* and cat_2 : *DET NOUN*. For the purpose of our analysis, we will consider (cat_1, cat_2) to be a minimal pair, as *the dog* can be considered a lexicalized instance of the more abstract construction *DET NOUN*. Using a distributional analysis, we can capture how the contexts of cat_1 and cat_2 vary, and how this variation is associated with generalization. If their cosine similarity decreases during training, it means that their contexts become more and more dissimilar: the network produces *DET NOUN* in new contexts which do not perfectly overlap with those of *the dog*, indicating that the network's babbling is becoming more productive (a graphical representation is given in Figure 5). In this case, we theorize that cat_2 has been recognised as a partially independent pattern from cat_1. If, on the contrary, their cosine similarity increases, we might deduce that the network has recognized cat_2 as partly unnecessary: it is correcting an overgeneralization.

We restrict this analysis to the CHILDES corpus. We build distributional vector spaces for the input and each stage of babbling using the DISSECT toolkit (Dinu et al., 2013). We consider catenae

---

[8]The complete set of correlation values is reported in supplementary material

[9]The Jaccard index between CHILDES and OpenSubtitles remains above 0.5, even when considering the top 1M catenae, while the same index computed between CHILDES and Simple Wikipedia drops to around 0.13.
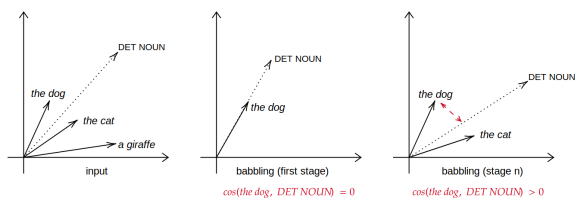
Figure 5: Let us assume that the input presents various lexicalized instances of the pattern *DET NOUN* (e.g. *the dog*, *the cat*, *a giraffe*). Our hypothesis is that the network will only be able to capture its more stereotypical instances (i.e., *the dog*), and the distributions of *the dog* and *DET NOUN* will thus almost perfectly overlap in the first stages of babbling (the length of vectors in the figure is just for exemplification). At later stages, the language produced by the network will show greater traits of productivity: the distribution of *DET NOUN* might show that its cosine distance to *the dog* has increased as it is now instantiated by two different lexicalized patterns (*the dog* and *the cat*) that are produced in dissimilar contexts.

composed by 2 or 3 elements as targets/contexts, and define co-occurrence as the presence of two catenae in the same sentence. Co-occurrences are weighted with PPMI and the space reduced to 300 dimensions with SVD. We then extract minimal pairs (cat_1, cat_2) of catenae from the input text, where cat_1 is an instance of cat_2. For each pair, we compute their cosine similarity in all distributional spaces, and the difference in cosine between the last and first babbling (see Table 4).

We then compute average distributional shifts and cosine similarities, grouping all pairs by cat_1 and cat_2 values (for instance, we average all pairs that show abstractions of cat_1 : *a minute*, as well as pairs that show instantiations of cat_2 : *DET NOUN*). Some averages are shown in Table 5.

We finally split catenae in three bins based on average distributional shift and investigate the influence of input similarity over the abstraction behaviour of a construction. Our hypothesis is that catenae that underwent the highest shifts during training were those showing intermediate levels of similarities in the input distributional space. Indeed, pairs with very high input similarities are unlikely to exhibit abstraction: according to constructionist intuition, their distributional similarity means that the catena that is part of the *Constructicon* is the least abstract one, and there is no need for the more abstract category. Low similarity pairs, on the other hand, may simply contain unrelated catenae.
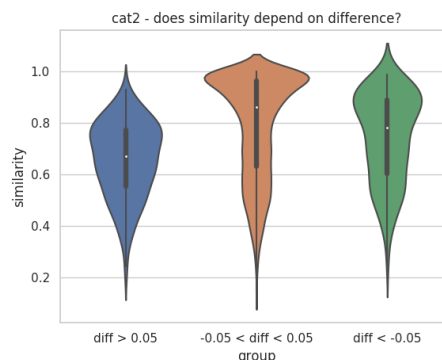
To test our hypothesis, we perform a Kruskall-

Figure 6: Distribution of average cosine similarities for the three groups of cat_2, showing low, intermediate and high average shifts respectively.

Wallis one-way analysis of variance test, that turn out to be significant for groupings made on both cat_1 and cat_2 lists.[10] The result is confirmed by Dunn's posthoc test. We show results for the test performed on the cat_2 list in Table 6 and Figure 6.

7 Discussion and future work

Usage-based computational accounts have already shown to be able to explain puzzling phenomena in acquisition (Freudenthal et al., 2015; McCauley and Christiansen, 2019) or to induce syntactic rules in an unsupervised manner (Solan et al., 2005), making use of surface properties of the language signal like transitional probabilities or basic distributional analysis. However, despite being rooted in the psychological literature and yielding fundamental psycholinguistic results, the models presented in such investigations are often not comparable to studies involving neural language models, as the former are usually less flexible and less scalable to large amounts of data than the latter.

In this paper, we have reviewed relevant work concerning the assessment of grammatical abilities in neural language models and noted the lack of variety in both the input data fed to ANNs (I) and the theoretical framework used in analysing the output language (Λ). In line with the existing usage-based computational accounts, we have introduced a methodology to evaluate the level of productivity of an LSTM trained on limited, child-directed data, using inspirations from constructionist approaches.

We have been able to show that neural networks approximate the distribution of constructions at a quite refined level when trained over a bare 3M

---

[10] $p = 6.988142426844016e-28$ for cat_1 and $p = 7.420868598608134e-32$ for cat_2

cat_1	cat_2	input	BM	5	10	15	20	25	30	35	distributional shift
a minute	a _NOUN	0.28	0.32	0.71	0.51	0.44	0.39	0.38	0.37	0.34	0.37
a minute	a @root	0.13	0.19	0.49	0.37	0.26	0.20	0.21	0.22	0.20	0.30
you _VERB it	_PRON @root @expl	0.10	0.19	0.46	0.28	0.25	0.25	0.19	0.17	0.21	0.25
you _VERB you	you _VERB @iobj	0.28	0.40	0.68	0.56	0.47	0.49	0.39	0.42	0.43	0.25
we can _VERB	_PRON can @root	0.51	0.54	0.79	0.74	0.59	0.54	0.55	0.61	0.57	0.22
go _VERB @obj	_VERB @conj @obj	0.64	0.72	0.56	0.74	0.70	0.74	0.72	0.72	0.72	-0.16
_AUX hungry	@cop @conj	0.68	0.52	0.36	0.39	0.44	0.45	0.47	0.42	0.59	-0.24
can get	can @advcl	0.55	0.54	0.24	0.36	0.45	0.48	0.43	0.39	0.52	-0.28

Table 4: Pairs of catenae (cat_1, cat_2), their cosine similarity in the space obtained from CHILDES, in the space obtained from the best model (BM) and in all the intermediate models. The last column shows the difference between cosine similarity at epoch 5 and cosine similarity at epoch 35.

cat_1	shift	cosine	cat_2	shift	cosine
@nsubj @root so	0.18	0.43	more @root	0.2	0.21
@nsubj only @root	0.18	0.41	_AUX know @obj	0.19	0.66
what @root @obj	0.18	0.39	@advmod tell	0.17	0.64
what @advmod _VERB	0.16	0.19	@aux know @obj	0.16	0.71
only @root	0.16	0.38	@advmod can _VERB	0.15	0.76
more @root	0.16	0.23	know @obj	0.15	0.62
@root it @xcomp	0.15	0.61	a _NOUN	0.13	0.52
@det minute	0.15	0.25	might @root	0.13	0.70
_PRON only @root	0.15	0.53	_PRON @root n't	0.12	0.53
_VERB _DET minute	0.15	0.33	@root that _VERB	0.12	0.65
_PRON @root so	0.14	0.54	_VERB 'll @ccomp	0.12	0.71
_DET minute	0.134	0.33	_VERB me @obl	0.12	0.76

Table 5: Catenae with highest average shifts.

	negative	none	positive
negative	-	6.83e-06	4.57e-05
none	0.000	-	4.15e-29
positive	0.000	4.15e-29	-

Table 6: Dunn posthoc test on the three groups of c_2, showing low (< -0.05), intermediate ($-0.05 < x < 0.05$) and high (< 0.05) average shifts respectively.

specific sets of structures.

Most importantly, the introduced methodology, despite being preliminary, presents a number of features that make our study fit in the usage-based theoretical framework while also using neural networks as language modeling tools, more specifically: (i) it posits no sharp distinction between lexicon and grammar: fully lexicalized, partially filled and purely syntactic patterns are all part of our constructicon and can play a similar role in production. Different items can therefore be represented compared, irrespective of their *lexical* nature; (ii) it makes no assumption about the stability of the constructicon: what is relevant for productivity at the earliest stages of learning might become superfluous later on; (iii) all items are seen as form-meaning pairs (i.e., constructions by definition, as in Goldberg, 2006): a novel way of modeling constructional meaning is therefore introduced and represents a promising path for future studies; (iv) distributional semantics is used both as a powerful quantitative tool and as a usage-based cognitive hypothesis, which leads us to specific assumptions about the cognitive format and origin of semantic representations (Lenci, 2008), and seems in line with the view of constructions as "invitations to form categories" (Goldberg, 2019).

Finally, we must account for potential biases introduced by applying dependency parsing to both input data and neural babbling: while this step is necessary to extract catenae, it introduces a non-negligible amount of noise, as the available pipelines are typically trained on different varieties than the ones considered in this study. In particular, the parser is somehow projecting its own categories, which have been acquired in a different setting and probably on a different variety, on our data. This currently limits the transferability of our results. Besides looking for ways to circumvent this issue, further work includes a comparison of our results with a wider choice of models.

words from the CHILDES corpus, reproducing the distribution of grammatical patterns even when they are not fully lexicalized. The analysis in § 5 indicates that the linguistic variety of OpenSubtitles is a potentially relevant benchmark to further investigate language acquisition, due to its similarity to the CHILDES data. In contrast, Simple Wikipedia has proved to be dissimilar to child-directed speech. This large difference should be taken into consideration when it comes to evaluating the grammatical abilities on the network: many of the studies cited in § 2 use models trained on Wikipedia or similar varieties, which may complicate the acquisition of generic grammatical phenomena heavily present in child-directed language. The analysis in § 6 further illustrated how we can follow paths of abstraction by putting our grammar formalism in a vector space. Additional investigations are of course needed to confirm our results. In particular, we would like to target the behavior of some

Acknowledgements

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on LearningRepresentations*.

Afra Alishahi, Grzegorz Chrupała, and Tal Linzen. 2019. Analyzing and interpreting neural networks for nlp: A report on the first blackboxnlp workshop. *Natural Language Engineering*, 25(4):543–557.

Geoff Bacon and Terry Regier. 2019. Does bert agree? evaluating knowledge of structure dependence through agreement relations. *arXiv preprint arXiv:1908.09892*.

Colin Bannard, Elena Lieven, and Michael Tomasello. 2009. Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, 106(41):17284–17289.

Marco Baroni. 2020. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307.

Helen L Bee, Lawrence F Van Egeren, Ann Pytkowicz Streissguth, Barry A Nyman, and Maxine S Leckie. 1969. Social class differences in maternal teaching strategies and speech patterns. *Developmental Psychology*, 1(6p1):726.

Noam Chomksy. 1959. Review of skinner's verbal behaviour. *Language*, 35:26–58.

Noam Chomsky. 1968. *Language and Mind*. New York: Harcourt Brace Jovanovich.

Noam Chomsky. 1981. Lectures on government and binding. *Dordrecht: Foris*.

Noam Chomsky. 1995. *The minimalist program*. MIT Press.

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.

Morten H Christiansen and Nick Chater. 2016. *Creating language: Integrating evolution, acquisition, and processing*. MIT Press.

Eve V Clark. 2009. *First language acquisition*. Cambridge University Press.

Hannah Cornish, Rick Dale, Simon Kirby, and Morten H Christiansen. 2017. Sequence memory constraints give rise to language-like structure through iterated learning. *PloS one*, 12(1).

Stephen Crain and Paul Pietroski. 2001. Nature, nurture and universal grammar. *Linguistics and philosophy*, 24(2):139–186.

Alejandrina Cristia, Emmanuel Dupoux, Michael Gurven, and Jonathan Stieglitz. 2019. Child-directed speech is infrequent in a forager-farmer population: a time allocation study. *Child development*, 90(3):759–773.

Tim Van de Cruys. 2011. Two multivariate generalizations of pointwise mutual information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 16–20. Association for Computational Linguistics.

Forrest Davis and Marten van Schijndel. 2020. Recurrent neural network language models always learn English-like relative clause attachment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1979–1990, Online. Association for Computational Linguistics.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013. DISSECT - DIStributional SEmantics composition toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 31–36, Sofia, Bulgaria. Association for Computational Linguistics.

Jonathan Dunn. 2017. Learnability and falsifiability of construction grammars. *Proceedings of the Linguistic Society of America*, 2:1–1.

Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Charles J Fillmore. 1988. The mechanisms of "construction grammar". In *Annual Meeting of the Berkeley Linguistics Society*, volume 14, pages 35–55.

Daniel Freudenthal, Julian M Pine, Gary Jones, and Fernand Gobet. 2015. Simulating the cross-linguistic pattern of optional infinitive errors in children's declaratives and wh-questions. *Cognition*, 143:61–76.

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.

Adele E Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Adele E Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Adele E Goldberg. 2019. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Kristina Gulordava, Piotr Bojanowski, Édouard 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, Volume 1 (Long Papers)*, pages 1195–1205.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Betty Hart and Todd R Risley. 1995. *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. 2002. The faculty of language: what is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

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

Thomas Hoffmann, Graeme Trousdale, and Ray Jackendoff. 2013. Constructions in the parallel architecture.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P Levy. 2020. A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Paul Kay and Charles J Fillmore. 1999. Grammatical constructions and linguistic generalizations: the what's x doing y? construction. *Language*, pages 1–33.

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, Melbourne, Australia. Association for Computational Linguistics.

Yair Lakretz, Cognitive Neuroimaging Unit, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in lstm language models. In *Proceedings of NAACL-HLT*, pages 11–20.

Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual review of Linguistics*, 4:151–171.

Michael A Lepori, Tal Linzen, and R Thomas McCoy. 2020. Representations of syntax [mask] useful: Effects of constituency and dependency structure in recursive lstms. *arXiv preprint arXiv:2005.00019*.

John D Lewis and Jeffrey L Elman. 2001. Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proceedings of the 26th annual Boston University conference on language development*, volume 1, pages 359–370. Citeseer.

David J Lewkowicz, Mark A Schmuckler, and Diane MJ Mangalindan. 2018. Learning of hierarchical serial patterns emerges in infancy. *Developmental psychobiology*, 60(3):243–255.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside bert's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253.

Tal Linzen and Marco Baroni. 2020. Syntactic structure from deep learning. *arXiv preprint arXiv:2004.10827*.

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.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Barbara Lust. 1999. Universal grammar: The strong continuity hypothesis in first language acquisition. *Handbook of Child Language Acquisition*.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. Third Edition*. Lawrence Erlbaum Associates.

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.

Danielle Matthews and Colin Bannard. 2010. Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child-directed speech. *Cognitive science*, 34(3):465–488.

Stewart M McCauley and Morten H Christiansen. 2019. Language learning as language use: A cross-linguistic model of child language development. *Psychological review*, 126(1):1.

James L McClelland. 1992. Can connectionist models discover the structure of natural language. *Minds, Brains and Computers*, pages 168–189.

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*.

R Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *arXiv preprint arXiv:2001.03632*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4034–4043.

Fernando Nogueira. 2014–. Bayesian Optimization: Open source constrained global optimization tool for Python.

Timothy Osborne. 2006. Beyond the constituent-a dependency grammar analysis of chains. *Folia Linguistica*, 39(3-4):251–297.

Timothy Osborne and Thomas Groß. 2012. Constructions are catenae: Construction grammar meets dependency grammar.

Timothy Osborne, Michael Putnam, and Thomas Groß. 2012. Catenae: Introducing a novel unit of syntactic analysis. *Syntax*, 15(4):354–396.

Timothy J Osborne. 2018. Tests for constituents: What they really reveal about the nature of syntactic structure. *Language Under Discussion*, 5(1):1–41.

Giulia Rambelli, Emmanuele Chersoni, Philippe Blache, Chu-Ren Huang, and Alessandro Lenci. 2019. Distributional semantics meets construction grammar. towards a unified usage-based model of grammar and meaning. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 110–120.

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 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.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. 2018. Ordered neurons: Integrating tree structures into recurrent neural networks. In *International Conference on Learning Representations*.

Zach Solan, David Horn, Eytan Ruppin, and Shimon Edelman. 2005. Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, 102(33):11629–11634.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019*.

Michael Tomasello. 2003. *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Ke M Tran, Arianna Bisazza, and Christof Monz. 2018. The importance of being recurrent for modeling hierarchical structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4731–4736.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2019. Blimp: A benchmark of linguistic minimal pairs for english. *arXiv preprint arXiv:1912.00582*.

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.