LANGUAGE, COGNITION AND NEUROSCIENCE
2022, VOL. 37, NO. 4, 420–439
https://doi.org/10.1080/23273798.2021.1980595

Routledge
Taylor & Francis Group

REGULAR ARTICLE

Hierarchy in language interpretation: evidence from behavioural experiments and computational modelling

Cas W. Coopmans [a,b], Helen de Hoop [b], Karthikeya Kaushik[c], Peter Hagoort [a,c] and Andrea E. Martin [a,c]

[a]Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands; [b]Centre for Language Studies, Radboud University, Nijmegen, Netherlands; [c]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands

ABSTRACT

It has long been recognised that phrases and sentences are organised hierarchically, but many computational models of language treat them as sequences of words without computing constituent structure. Against this background, we conducted two experiments which showed that participants interpret ambiguous noun phrases, such as *second blue ball*, in terms of their abstract hierarchical structure rather than their linear surface order. When a neural network model was tested on this task, it could simulate such "hierarchical" behaviour. However, when we changed the training data such that they were not entirely unambiguous anymore, the model stopped generalising in a human-like way. It did not systematically generalise to novel items, and when it was trained on ambiguous trials, it strongly favoured the linear interpretation. We argue that these models should be endowed with a bias to make generalisations over hierarchical structure in order to be cognitively adequate models of human language.

ARTICLE HISTORY
Received 26 March 2021
Accepted 8 September 2021

KEYWORDS
Syntax; constituency; meaning; human-like generalisation; LSTM

1. Introduction

The ability to use language is a hallmark of the human mind. The formal structures of human language reveal the wealth of representational infrastructure that our brains deploy to guide our linguistic behaviour. As such, even in a short phrase like *these two blue balls* lies a hidden signal about how the mind structures information. For this simple four-word phrase, there are 24 logically possible word orders, yet only 14 of these are attested in the world's languages (Cinque, 2005). Strikingly, the word order in English and its mirror variant (*balls blue two these*) are by far the most frequent (Cinque, 2005; Greenberg, 1963), reflecting the selection of word orders that transparently map to the hierarchical structure of the noun phrase (Culbertson & Adger, 2014; Martin et al., 2020). The word "hierarchical" here refers to the representational format of constituent structure: words are embedded into constituents, which are in turn recursively embedded into larger constituents, creating hierarchically organised syntactic structures which are often visually denoted by means of tree structures (see Figure 1(a)). It has long been argued that the semantic interpretation of phrases and sentences is linked to this hierarchical constituent structure (e.g. Chomsky, 1957; Everaert et al., 2015; Heim & Kratzer, 1998; Jackendoff, 1972; Partee, 1975; Pinker, 1999). That is, syntactic operations are defined over hierarchical structure rather than linear order (i.e. they are structure-dependent; Chomsky, 1957), and semantic dependencies (like scope, the fact that *two* applies to *blue balls* rather than *balls* alone[1]) directly follow from such hierarchically organised constituent structure.

Despite these arguments in theoretical linguistics, however, an alternative view holds that language use can be accounted for in terms of sequential rather than hierarchical structure (e.g. Bybee, 2002; Christiansen & Chater, 2015; Frank et al., 2012). A core aspect of this view, which has been championed by several authors in different proposals, is that constituency is not a basic structure but rather an epiphenomenon, emerging from frequently occurring sequential patterns in language, which are "chunked" into sequences without much internal structure (Bybee, 2002; Christiansen & Chater, 2015; Frank et al., 2012). In other words, while this "linearity view" does not entail that hierarchical structure does not exist, it holds that hierarchy is not fundamental in language use. This view is strengthened by the recent successes of mainstream models in natural language processing (NLP), which treat sentences as linear strings of words. These models achieve

CONTACT Cas W. Coopmans cas.coopmans@mpi.nl Max Planck Institute for Psycholinguistics, Wundtlaan 1, P.O. Box 310, Nijmegen 6525 XD, Netherlands, Centre for Language Studies, Radboud University, P.O. Box 9103, Nijmegen 6500 HD, Netherlands

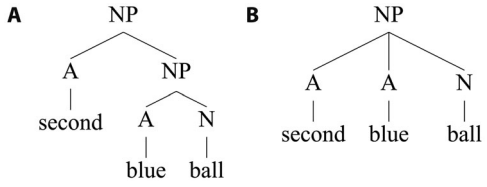

Figure 1. Hierarchical (a) and linear (b) representations for the phrase *second blue ball*.

remarkably good performance, arriving at around 93% accuracy on several diagnostics (e.g. Devlin et al., 2018), and are often used to account for behavioural data in psycholinguistic experiments (e.g. Christiansen & MacDonald, 2009; Frank & Bod, 2011; Gulordava et al., 2018; Linzen et al., 2016).

Against this background, we use the interpretation of ambiguous noun phrases such as *second blue ball* as a test of the idea that constituency is not fundamental in language use. We first show in two behavioural experiments that the interpretation of these phrases is based on their hierarchical rather than their linear structure, indicating that language interpretation can in fact be biased towards hierarchical constituency. We then train and test a recurrent neural network model on our task in order to see whether it is able to reproduce such "hierarchical" behaviour. In several simulations, we evaluate whether the model generalises in a human-like way. We show that it can simulate hierarchical behaviour, but only if the training data are unambiguously hierarchical. When it is trained on ambiguous data which are equally consistent with the linear and the hierarchical interpretation of *second blue ball*, it strongly favours the linear interpretation. Moreover, the model does not systematically generalise to items that were not observed during training. Overall, this leads us to conclude that without a predisposition for hierarchical structure, the model is not a cognitively adequate model of human language (Dehaene et al., 2015; Fitch, 2014).

1.1. Behavioural evidence for hierarchical structure

Broadly speaking, two kinds of evidence support the claim that words, phrases and clauses have internal hierarchical structure. First, syntactic operations, such as movement, deletion and substitution target constituents rather than individual words. These operations are said to be structure-dependent, and behavioural experiments have shown that children obey structure dependence as soon as they can be tested (e.g. Crain & Nakayama, 1987). Second, structure provides the unit of semantic interpretation, as can be seen in the structural ambiguity of words (e.g. *uninstallable*), phrases (e.g. *deep blue sea*) and clauses (e.g. *she saw the man with binoculars*), as well as the structure-dependent interpretation of anaphora, disjunction, negative polarity items, and other scope phenomena (Reinhart, 1983; see Crain et al., 2017 for a recent overview of the empirical data from acquisition). These facts about language structure show that constituents behave as units, both to syntactic operations and to semantic interpretation.

Furthermore, a large body of experimental evidence converges in showing how hierarchical structure explains language behaviour. Of particular relevance to the current study are three behavioural paradigms which investigate noun phrase interpretation. First, Lidz and colleagues used a preferential looking paradigm to show that 18-month-old infants interpret the pronominal *one* in *Look! A yellow bottle. Do you see another one?* as anaphoric with the constituent *yellow bottle* rather than with the bare noun *bottle*, consistent with the interpretation of anaphoric *one* in adult language (Lidz et al., 2003). Second, a cross-domain structural priming study by Scheepers and Sturt (2014) showed that people find adjective-noun-noun compounds more acceptable when their structure is congruent with a mathematical equation that they have solved just before. In their study, left-branching phrases, such as *organic coffee dealer* (i.e. [[organic coffee] dealer]), received higher ratings after left-branching equations (e.g. $25 \times 4 - 3$) than after right-branching equations (e.g. $25 - 4 \times 3$). Third, Culbertson and Adger (2014) exposed English learners of an artificial language to different noun phrases with only one postnominal modifier (i.e. N-Dem, N-Num, N-Adj), based on which they had to infer the relative ordering of the modifiers in a complex noun phrase (see also Martin et al., 2020). The training data were equally consistent with two possible grammars, one of which was similar to English in terms of the linear ordering of the modifiers (i.e. balls *these two blue*), while the other was similar to English in terms of the abstract structure of the noun phrase (i.e. [[[[balls] *blue*] *two*] *these*]). The learners consistently favoured the order that was structurally similar to English, despite its dissimilarity to English in terms of surface statistics. In line with this finding, a recent study on artificial rule learning showed that people from different age groups and different cultural and educational backgrounds spontaneously infer and generalise abstract hierarchical structure after exposure to sequences whose structure is fully consistent with both hierarchical rules (based on recursive center-embedding) and linear rules (based on ordinal position; Ferrigno et al., 2020). Combined, these studies

demonstrate that people represent noun phrases as hierarchical structures rather than as linear sequences. Moreover, the studies by Lidz et al. (2003) and by Culbertson and Adger (2014) indicate that this hierarchical bias does not come from the environment but rather reflects an inherent property of the linguistic system, which might also be present in other domains of cognition (Dehaene et al., 2015; Ferrigno et al., 2020; Fitch, 2014).

Evidence from the spontaneous creation of languages in language-deprived populations supports this latter point. Deaf children who are born to speaking parents and are not exposed to sign language in infancy spontaneously develop a gestural system for communication (Goldin-Meadow, 2003). This system, called *homesign*, has many of the properties of natural language, including hierarchically organised levels of recursive constituent structure and structure-dependent operations, such as substitution (Goldin-Meadow, 2003; Hunsicker & Goldin-Meadow, 2012). For example, in homesign, multi-gesture combinations that refer to a single nominal entity (e.g. a demonstrative gesture and a noun gesture: "that bird") function both syntactically and semantically like single-gesture nominals. They can substitute for a single noun ("bird") and can be embedded in a hierarchically structured clause, to yield a signed clause with the hierarchical structure [[that bird] pedals] rather than the flat structure [that bird pedals] (Hunsicker & Goldin-Meadow, 2012). Because the multi-gesture nominals produced by homesigners are effectively absent in the gestures of their hearing family members, they reveal that the homesigners themselves are the source of these structural properties in their linguistic system (Flaherty et al., 2021).

1.2. Computational modelling of hierarchical structure

While the behavioural evidence strongly supports the hierarchical view, the linearity view is strengthened by recent results from computational studies of language acquisition. Most contemporary language models are not endowed with a cognitive architecture that supports the acquisition and knowledge of linguistic information (e.g. hierarchical representations, structure dependence, or compositionality), yet they perform quite well on a range of language tasks. In particular, recent computational research with recurrent neural network (RNN) models has shown that these models often perform quite accurately on tasks which are thought to require knowledge of hierarchical structure, such as subject-verb agreement and question formation. For example, RNNs can learn to generate the correct agreement in long-distance dependencies (e.g. *The boy who likes the girls has …*) and to move the right verb in constructing complex yes-no questions (e.g. *Has the boy who likes the girls … ?*), seemingly without invoking hierarchical structure (Gulordava et al., 2018; Linzen et al., 2016; McCoy et al., 2018, 2020; Tran et al., 2018). Moreover, RNNs are able to generalise very well to novel grammatical constructions when these feature a mixture of examples that were observed in the training set, but they fail to systematically generalise across items in the training set to compose novel items (Baroni, 2020; Lake & Baroni, 2018; Loula et al., 2018). These findings show that RNN models show impressive generalisation ability, apparently without relying on systematic compositionality.

It is often the case that the data on which these models are trained is both qualitatively and quantitatively very different from the linguistic input children receive (Linzen, 2020; Linzen & Baroni, 2021). Recent studies have sought to address this issue by exposing the model during training only to ambiguous data, from which multiple generalisations are possible (e.g. McCoy et al., 2018, 2020; Mulligan et al., 2021). During the test phase, the model is then evaluated on items for which these generalisations make different predictions. The idea behind this training-test regime is that the model's performance on test trials reveals its specific inductive biases. Comparing this performance to human behaviour in the experimental paradigms discussed above (Culbertson & Adger, 2014; Ferrigno et al., 2020; Martin et al., 2020), we can evaluate whether these models generalise in a human-like way. Initial results from these studies show that some RNN architectures can make human-like syntactic generalisations, in particular when the training data contain cues to hierarchical structure (McCoy et al., 2018).

In short, while most computational language models do not explicitly incorporate structure dependence, they appear extremely proficient in a range of complex language tasks if they are trained on quantitatively and qualitative rich data. This reveals a possible gap between the validity of these models as models of human cognition and their ability to achieve human-like behaviour in certain circumstances. We approach this issue by comparing the performance of a long short-term memory (LSTM) neural network to the behaviour of human participants on a task that requires hierarchically structured knowledge. The following sections first describe the task and results from the behavioural experiments.

1.3. Background of the present study

In two experiments, we tested whether people interpret ambiguous noun phrases such as *second blue ball* as a

hierarchical structure or as a linear string. On the hierarchical interpretation, which is derived from the right-branching structure depicted in Figure 1(a), the structure encodes semantic scope. The ordinal *second* takes scope over the constituent *blue ball*, and the whole refers to the second among blue balls. On the linear interpretation, instead, *second* and *blue* are interpreted conjunctively, and they independently modify the noun *ball* (i.e. the ball that is blue *and* second). Here, the conjunctive (linear) interpretation is associated with the flat representation depicted in Figure 1(b). However, we note that this is not the only possible way in which that interpretation can be represented. It could also be derived from a hierarchical structure, for instance by means of a conjunction phrase which first combines *second* and *blue*, and is then combined with *ball*. In contrast, the scopal (hierarchical) interpretation of *second blue ball* can only be derived from a nested constituent structure (i.e. Figure 1(a)). Because the hierarchical interpretation cannot be derived without hierarchical structure (as in the linear representation in Figure 1 (b)), consistently hierarchical responses should be taken as evidence against the view that hierarchical structure is unnecessary to account for language interpretation.

To show how the semantics corresponding to these phrases relates to their structure (Partee, 2007; Spenader & Blutner, 2007), we provide the lambda expressions for the noun *ball* (which is of type <e,t>), the intersective adjective *blue* and the adjective *second* (which are both predicate modifiers of type <<e,t>,<e,t>>) below:

1. *ball*: λx[ball(x)]
2. *blue*: λPλx[P(x) & blue(x)]
3. *second*: λPλx[P(x) & ∃!y[P(y) & y < x]]
   where < indicates a type of ordering relationship (i.e. y precedes x on some dimension, such as space or time).

In these expressions, *P* refers to a one-place predicate, i.e. a set of individuals that is the denotation of a noun such as *ball*. Hence, we get the following lambda expressions that correspond to the noun phrases *blue ball* and *second ball*, which are both of type < e,t>:

4. *blue ball*: λx[ball(x) & blue(x)]
5. *second ball*: λx[ball(x) & ∃!y[ball(y) & y < x]]

Combining these expressions yields the hierarchical right-branching interpretation of the complex noun phrase *second blue ball* (corresponding to Figure 1(a)), as expressed in (6):

6. Hierarchical interpretation: λx[ball(x) & blue(x) & ∃!y [ball(y) & blue(y) & y < x]]

This means that *second blue ball* on the hierarchical interpretation refers to the set of elements x that are a member of the intersection of the set of balls and the set of blue things, such that there is exactly one other element in this intersection, which is the set of blue balls, preceding x (in one way or another). Clearly, in this interpretation, *second* applies to the set of blue balls, which means that *blue* and *ball* are combined to form a constituent that serves as the argument of *second*.

On the linear interpretation of *second blue ball* this would not be the case. Here, *second blue ball* would denote the set of elements x that are a member of the intersection of the set of balls and the set of blue things, such that there is exactly one other element in the set of balls preceding x. On this interpretation, the phrase refers to the second ball, which is blue (i.e. a green ball was in the first position). The lambda expression for the linear interpretation of *second blue ball* (corresponding to Figure 1(b)) is given in (7):

7. Linear interpretation: λx[ball(x) & blue(x) & ∃!y[ball(y) & y < x]]

While these two interpretations could yield the same referent (Figure 2(a)), this need not be the case: based on the context in which *second blue ball* is presented, the linear and hierarchical interpretations can diverge (Figure 2(b)). This divergence forms the basis of the current study.

The idea was based on a set of acquisition experiments conducted in the 1980s, in which it was investigated how children acquire and interpret prenominal modifier sequences (Hamburger & Crain, 1984; Matthei, 1982). Matthei (1982) asked five-year old children to point to the *second blue ball* in an array of coloured balls in which the linear and hierarchical interpretations yielded a different answer (Figure 2(b)). The children interpreted the phrase intersectively, pointing to the ball that was blue and in the second position, rather than to the second among blue balls. This was taken to indicate that the children had built an unembedded, linear representation. In a reply to this study, Hamburger and Crain (1984) noted that Matthei's (1982) results reflected the children's inability to deal with the cognitive complexity of the task, which might have concealed their hierarchical grammatical knowledge. They attempted to reduce the nature of the planning component underlying these linguistic expressions by letting children first *point to the first blue ball*, and then

Figure 2. Example arrays for the target *second blue ball*, corresponding to a convergent (a) and a divergent (b) trial.

point to the second one. The children's interpretations of *one* in this scenario are indicative of whether they relied on a linear representation of *first blue ball*, in which case *one* can only refer to *ball*, or on a hierarchical representation, in which *one* can also refer to *blue ball*. Similar to the infants in the Lidz et al. (2003) study, four-year-old children took *one* as anaphoric with the constituent *blue ball*, indicating that they relied on a hierarchical representation of *first blue ball*. We adopted a similar experimental paradigm, but chose to use full noun phrases rather than anaphoric pro-forms, given the debate about whether *one* indeed substitutes for syntactic constituents (Goldberg & Michaelis, 2017; Payne et al., 2013).

2. Methods and results

2.1. Experiment 1

The first experiment is a replication of the original study by Matthei (1982), but with only adults. 20 native speakers of Dutch (14 females, mean age = 21.9 years, range = 19–27 years) participated in the experiment, none of whom were colour-blind. All participants gave written informed consent to take part in the experiment, which was approved by the Ethics Committee of the Faculty of Social Sciences at Radboud University Nijmegen. The experiment was conducted in Dutch, but for ease of exposition, the stimuli are translated here into English, which in these sentences has the same surface word order as Dutch. Participants had to click on a target denoted by a noun phrase containing an ordinal, a colour adjective and a noun referring to the shape of the target, such as *second blue ball*. Two example arrays, corresponding to the two conditions, are presented in Figure 2.

In the convergent condition, the hierarchical (non-intersective) and linear (intersective) interpretation converge on the same item. For example, the second blue ball in Figure 2(a) is both the second among blue balls (hierarchical) and also the ball that is blue and in second position (linear). In the divergent condition, the linear and hierarchical interpretation yield a different answer. While the second ball in the array in Figure 2 (b) is blue (linear), it is not the second among blue balls, which is in fourth position (hierarchical).

The convergent condition was not present in the original studies (Hamburger & Crain, 1984; Matthei, 1982). The responses in this condition do not dissociate between hierarchical and linear interpretations, and serve as fillers to reduce the potential influence of pragmatic factors. That is, one could argue that participants only give hierarchical answers in response to *second blue ball* on divergent trials because they take the mere presence of *blue* to indicate that they should not interpret the phrase as referring to the second ball. Had that been the intended target (e.g. in the picture of Figure 2(b)), then it could have been referred to as *second ball*, thus making the addition of *blue* redundant and therefore pragmatically odd.[2] By making sure that half of the trials contains a redundant colour adjective, we intended to make participants less sensitive to the effect of redundancy on interpretation, thereby making it less likely that their behaviour on divergent trials would be driven by pragmatic factors.

Each trial consisted of the written sentence "Click on the [target]" and an array of eight blue or green balls, visually presented at the same time on a computer screen. The target was always described using an ordinal, a colour adjective, and the noun *ball*. The ordinals first, second, third, fourth, fifth, and sixth were used. There were 192 trials, half of which were divergent, the other half were convergent. In both conditions, all ordinals were used 16 times in the target phrase, and they were equally often combined with green as with blue. Convergent trials were created as follows: all items to the left of the target were the same as the target, and all items to the right were randomised. For the divergent trials, there were two possible targets: a linear one and a hierarchical one.[3] For every ordinal, the position of the hierarchical target was randomly chosen among the positions to the right of the linear target. The positions to the left of the hierarchical target were then filled with the right number of items that were the same as the target. For instance, for the target *sixth green ball*, the hierarchical target could be in the positions 7 or 8. To the left of this position five green balls were placed, and one of these green balls was in sixth position (linear target). The other positions are filled with blue balls. Correct answers on convergent trials were coded as hierarchical/linear, while all other items were coded as error. On divergent trials, answers were coded as hierarchical, linear, or error.

2.1.1. Results

The results of experiment 1 are presented in Figure 3. The graph on the right contains the results for divergent

trials, which shows that of all correctly answered trials, participants gave a hierarchical answer 99.8% of the time. Only three answers were according to a linear interpretation. To test this effect, we applied a logistic regression model in R (R Core Team, 2020) with only an intercept to the binary output variable (hierarchical vs. linear), which showed that participants gave more hierarchical than linear answers, $\beta = -6.27$, $SE = 0.58$, $Wald\ z = -10.84$, $p < .001$.

While these results strongly suggest that the participants used hierarchical syntax, there is one alternative interpretation that does not need to rely on constituent structure. In this interpretation, *second* applies to the set of *blue* things first, hence forming a complex adjective *second blue*, which is then applied to the noun *ball* (e.g. a ball that is second among blue items). This is similar to phrases in which second modifies an adjective, e.g. *second biggest ball* (which is the ball that is the second biggest, but not necessarily the second ball), and phrases in which blue is modified by an adverb, e.g. *very blue ball* (which is very blue, not very ball). Because the arrays of items contained only balls, this approach always yields the same target as the hierarchical interpretation.

Importantly, while this alternative interpretation can be represented in a constituent structure (as in the left-branching structure in Figure 4(b)), it does not, strictly speaking, need hierarchy. In the right-branching "hierarchical" structure in Figure 4(a), a relationship is established between the element *second* and a constituent (i.e. a constituent is modified). Such a constituency-based relationship is not needed to represent the meaning of the left-branching structure in Figure 4(b), which would be expressed as follows:

8. Left-branching interpretation: λx[ball(x) & blue(x) & ∃!y[blue(y) & y < x]]

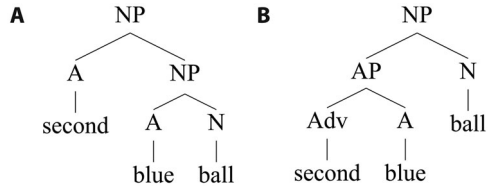

Figure 4. Right-branching (a) and left-branching (b) representations for the phrase *second blue ball*.

On this interpretation participants would choose the second blue thing in a sequence, which happens to be a ball (e.g. when the first position contains a blue triangle). While right-branching interpretations must rely on constituency, left-branching interpretations can, but need not do so. A second experiment was undertaken to adjudicate between the right-branching and left-branching interpretation.

2.2. Experiment 2

20 native speakers of Dutch (15 females, mean age = 23.0 years, range = 18–28 years) took part in the experiment after their written informed consent was obtained. None of the participants were colour-blind or had participated in experiment 1. The experiment was almost identical to experiment 1, except that the set of items in the array also contained blue and green triangles. As there were now two shapes, the noun provided crucial information for the identification of the target. Each trial contained two potential targets. For the target *second blue ball*, the "right-branching" interpretation, corresponding to the right-branching structure in Figure 4(a), again refers to the second among blue balls (fifth item in Figure 5). The other interpretation, which could be represented in a left-branching structure (Figure 4(b)), refers to the second blue item, which is a

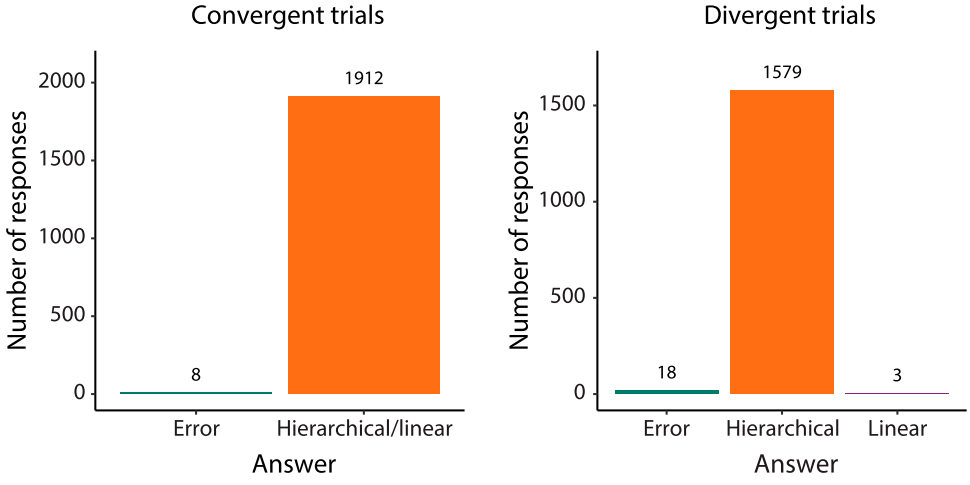

Figure 3. Responses in the convergent and divergent conditions of experiment 1.

Figure 5. Example array for the target *second blue ball*. The left-branching target is in third position, while the right-branching target is in fifth position.

ball (third item in Figure 5). The right-branching and left-branching interpretations were both always available, but never converged on the same item.

There were again 192 trials. All ordinals were used 32 times in the target phrase. For each ordinal, the target was equally often a blue ball, a blue triangle, a green ball, and a green triangle. We made sure that the left-branching and right-branching interpretations never converged on the same item by placing one item with the same colour but a different shape as the target at a random position to the left of the left-branching target. In Figure 5, this is the blue triangle on the left, which makes the leftmost blue ball the second blue item (the left-branching target). The presence of a blue triangle does not affect the position of the right-branching target, which is the second among blue balls.[4]

2.2.1. Results

The results of experiment 2 are presented in Figure 6. Of all correctly answered trials, participants gave a right-branching answer 99.8% of the time. Only five answers were coded as left-branching answer. A logistic regression analysis of output type (right-branching vs. left-branching) showed that participants gave more right-branching than left-branching answers, $\beta = -6.45$, $SE = 0.45$, $Wald\ z = -14.42$, $p < .001$. These findings can only be captured using constituent structure, and therefore provide strong experimental evidence for the importance of hierarchical structure for semantic interpretation.

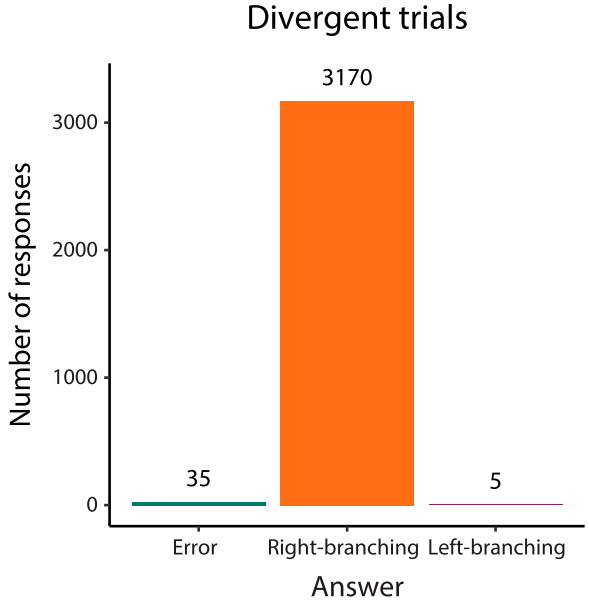

Figure 6. Responses in experiment 2.

2.3. Computational modelling

2.3.1. Methods

In order to test whether a computational model would show the same bias towards the hierarchical interpretation as the participants did, we trained and tested a state-of-the-art RNN model with a long short-term memory (LSTM) architecture (Hochreiter & Schmidhuber, 1997) on the task of Experiment 1.[5] The LSTM model, which was implemented with Keras (Chollet et al., 2015), had a many-to-one architecture, which is visually represented in Figure 7. The input to the model consisted of four one-hot vectors, sequentially presented in four timesteps. Recurrence is indicated by the fact that the model's current state is a function of its previous state (i.e. $a^{<t-1>}$) in combination with the input at the current timestep (i.e. x). The input vectors represent respectively the ordinal, colour, and shape of the target, as well as the picture. Each input vector had a length of 57, where the first 9 elements were reserved for words in the phrase (elements 1–6 represented the ordinals second through seventh, 7 and 8 represented the colours blue and green, and 9 represented the shape ball[6]) and the last 48 elements were reserved for

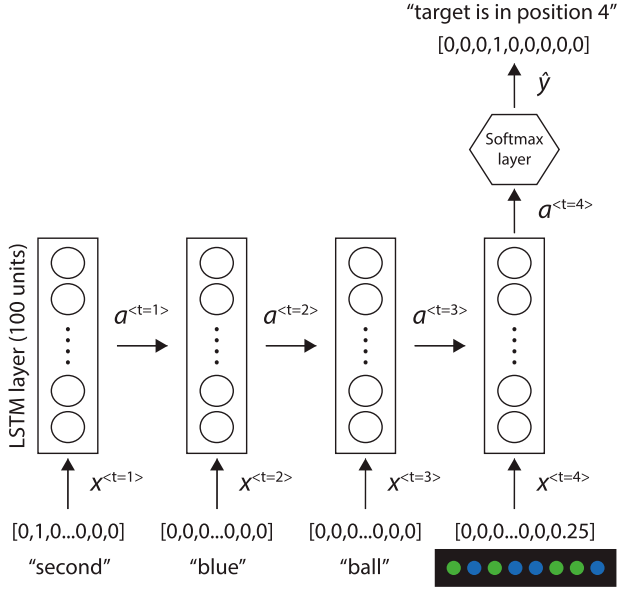

Figure 7. Visual representation of a trial for the LSTM, where x represents the input at timestep t and $a^{<t-1>}$ the activation state of the model after the previous timestep.

the eight-element picture, wherein each element had one of two colours and the shape ball (i.e. we need three bits to represent each feature). As a result, each picture vector would have 16 ones, so we normalised it to make sure that its net content is 1, in line with the other one-hot vectors. To give an example of an input vector, the word "blue" was represented as a 57-element vector which has a one in position 7 and zeros everywhere else.

The hidden layer consisted of 100 units, whose activation function at the last timestep was forwarded to a softmax layer, which provided the output of the network. The output was a nine-element one-hot vector which had a one at the position of the target (positions 1–8) on target-present trials or a one at position 9 to indicate that the target was absent from the picture. In short, the task of the model was to take the words and picture sequentially as input, and provide as output the position of the target.

The LSTM was trained in a supervised manner on datasets of different sizes (100-1000 trials, depending on the training set), in 50 epochs (100 steps per epoch) using the optimiser "Adam" (optimisation using stochastic gradient descent with a learning rate of 0.001) and the categorical-crossentropy loss function. For each dataset, the model was evaluated on 100 test trials, and this training-test evaluation was simulated 100 times.

2.3.1.1. Training and test datasets. We trained the LSTM on four different, artificially created datasets. In half of the trials in all datasets the target was present, in the other half, the target was absent.[7] While target-absent trials were not included in the behavioural experiments, we did include them in the datasets for the network because this ensures that the network cannot succeed by only paying attention to the ordinal. In all datasets, the training and test trials were mutually exclusive, never containing identical trials. Figure 8 presents a visual overview of the different training/test trials.

In the "linear" training and test set, the linear interpretation was present on target-present trials, and absent on target-absent trials. Moreover, on both target-present and target-absent trials, the hierarchical interpretation was also present, but the output showed that the training data were unambiguously about the linear interpretation, because trials were always divergent (cf. Figure 2(b)). To give an example, if the target was *second blue ball*, then the second ball in target-present pictures was blue, but it was not the second among the blue balls (i.e. the first ball was green; see Figure 8). Here it becomes clear why we included target-absent trials. If the target were always present, there would be a perfect statistical relationship between the ordinal and the output (i.e. *second blue ball* would always lead to target position 2). This could serve as a context-independent statistical heuristic for the model, as it would not need to incorporate information about the colour or shape of the target, or about the elements in the picture. By including target-present trials, we made sure that the model could not succeed by relying only on the information provided by the ordinal.

The "hierarchical" training and test set consisted of target-present trials in which the hierarchical interpretation was present and target-absent trials in which it was absent. All target-present trials were divergent (cf. Figure 2(b)), so the linear interpretation of the phrase would also be present, but the output was only in line with the hierarchical interpretation. On target-absent trials the hierarchical interpretation was absent but the linear interpretation was still present. For example, if the target phrase was *second blue ball*, then the second ball was blue on both target-present and target-absent trials, but it would not be the second among blue balls (in fact, on target-absent trials the second ball would be the only blue ball; see Figure 8).

The "ambiguous" training set was fully ambiguous between the hierarchical and linear interpretations of the target phrase, both on target-present and target-

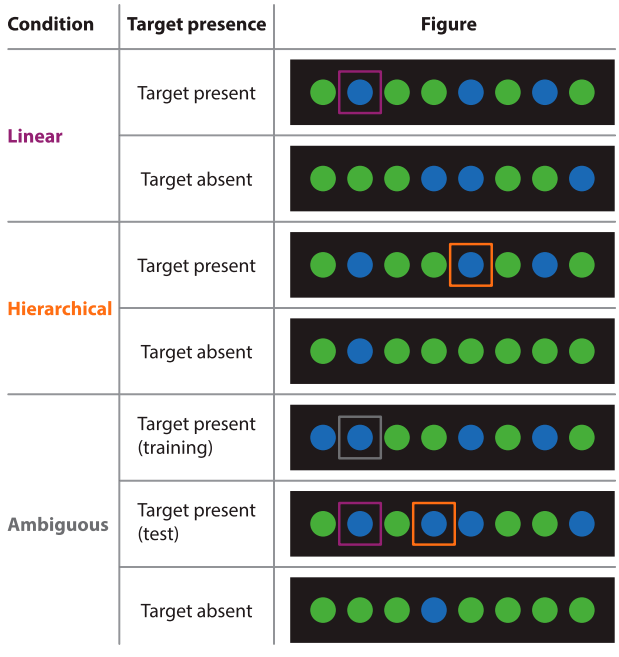

Figure 8. Examples of the different training/test trials in the computational simulations, ordered by condition and target presence. The target phrase for these trials is *second blue ball*. The squares in target-present trials indicate the target for each trial.

absent trials. While target-present training trials were always convergent (cf. Figure 2(a)), target-present test trials were always divergent (cf. Figure 2(b)). The model's answers on these test trials are thus informative about what the model has induced from ambiguous training data. On target-absent training and test trials, neither the linear nor the hierarchical interpretation was present. The ambiguous training set had only 100 trials. This has to do with the fact that target-present training trials are always convergent and thus limited in number, and that the number of unique trials varies per ordinal (e.g. for *seventh blue ball*, there are only two different target-present pictures (one in which the eighth ball is also blue, and one in which it is green), but for *second blue ball* there are 64 different target-present pictures). To make sure that the training and test sets contain roughly the same number of all ordinals, they were both fixed at a size of 100 trials.

The "mixed" training set contained both ambiguous and unambiguously hierarchical training trials. While the only possible generalisation from these data is the hierarchical interpretation, the linear interpretation is compatible with some of the trials. By varying the percentage of ambiguous trials (and thus the ratio between ambiguous and hierarchical trials), we examined how much unambiguously hierarchical data the model needs in order to consistently give hierarchical responses on test trials. The test trials were the same as those used after ambiguous training (i.e. divergent trials).

2.3.1.2. Generalisation to novel items. To further investigate what the model has learned after the hierarchical training regime, we tested its ability to generalise to items that were not seen during training. Specifically, we looked at the model's response to phrases that included the word "red" when the training data did not contain red at all (extrapolation), or only in combination with specific ordinals (interpolation). First, we trained the model on all items (green and blue balls), and then tested it on phrases with the word "third red ball" and pictures which included red balls. This type of generalisation is an instance of extrapolation, because the input contains features (i.e. the word "red", as well as red balls) that were not observed during training and therefore lie outside the training space (Marcus, 1998). Second, we tested the model's ability to interpolate, i.e. to generalise to an item that is composed of known features, and therefore lies within the training space (e.g. Baroni, 2020; Lake & Baroni, 2018). The model was trained on all combinations of features, including the colour "red", except the item "third red ball" (e.g. "second blue/green/red ball", "third blue/green ball", and pictures which included red balls). It was then tested on "third red ball". Here, the training data contains the distributional evidence that "red" and both "blue" and "green" pattern identically, and it contains information about how "third *x*" should be interpreted. Given that "third", "red" and "ball" have all been presented during training, the training data span a distribution that captures "third red ball", even though the combination of these items is new. Given that the new item lies within the parameter space, interpolation can be approached through linear regression. We therefore hypothesise that the model is able to interpolate from known data points to "third red ball". In order to see how well the model extrapolates and interpolates, we simulated each generalisation test 100 times. Because the hierarchical model in the main experiment reached over 90% accuracy after 500 training trials (discussed in Section 2.3.2 Results, Figure 10(b)), we trained the model in each simulation on 500 trials. As in the main experiment, it was evaluated on 100 test trials.

As reported in the results, the model was not able to systematically generalise its "hierarchical" knowledge to novel items, such as "third red ball". While the training data for the interpolation test contained the information that "red" functions the same as both "blue" and "green", it is possible that this distributional information was not sufficient to indicate the relatedness between these words. That is, there is no intrinsic relationship between the one-hot vectors [0, 1 … 0, 0] and [0, 0 … 1, 0], although they should be dependent if they are to represent the related words "red" and "blue". In an attempt to test the model's generalisation ability when it receives input vectors that are closely related, we used pre-trained word embeddings from Google's word2vec (Mikolov et al., 2013), which have been shown to capture the similarity between related words. The similarity between two multidimensional word embedding vectors can expressed in terms of the cosine of the angle between them. The closer this "cosine similarity" value is to 1, the smaller the angle between the vectors and thus the more similar the vectors (see the similarity matrix in Figure 9(a)).

As these word embeddings are 300-dimensional vectors, however, they might lead to overfitting given the limited size and scale of the training data. The model might overcapitalise on redundant aspects these big vectors, disabling them from dealing with novel input. We therefore used a dimensionality reduction technique based on Principal Component Analysis to reduce the size of the word embeddings to 10 (Shlens, 2014), in line with the size of our vocabulary.[8] This reduces the size of the vectors by maximising the

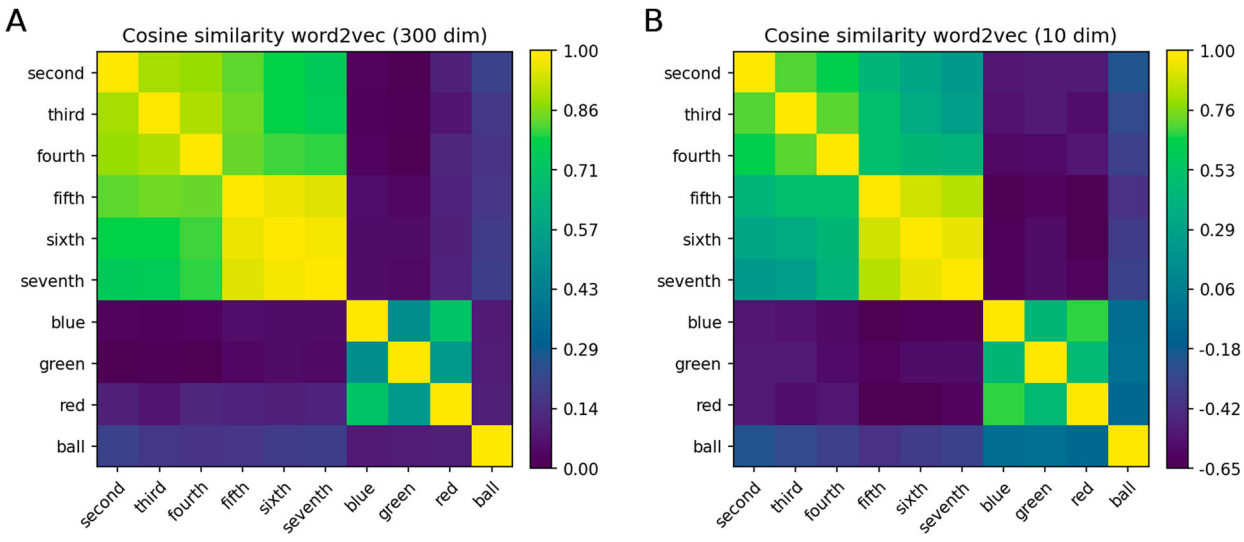

Figure 9. Heatmap of the cosine similarity between all 300-dimensionsal word embeddings (a) and between all 10-dimensional word embeddings (b). Note that in both cases the word embeddings capture the similarity between "blue", "green", and "red", as indicated by a large and positive cosine similarity.

variance between them, while retaining the essence of the original vectors. For our purpose it is important that the similarity between the colour words, which is the property over which generalisation is evaluated, is retained after dimensionality reduction (Figure 9(b)).

We repeated the two generalisation tests described above (including separate train-and-test evaluations) with both the full 300-dimensional word embeddings as well as the reduced 10-dimensional embedding vectors.

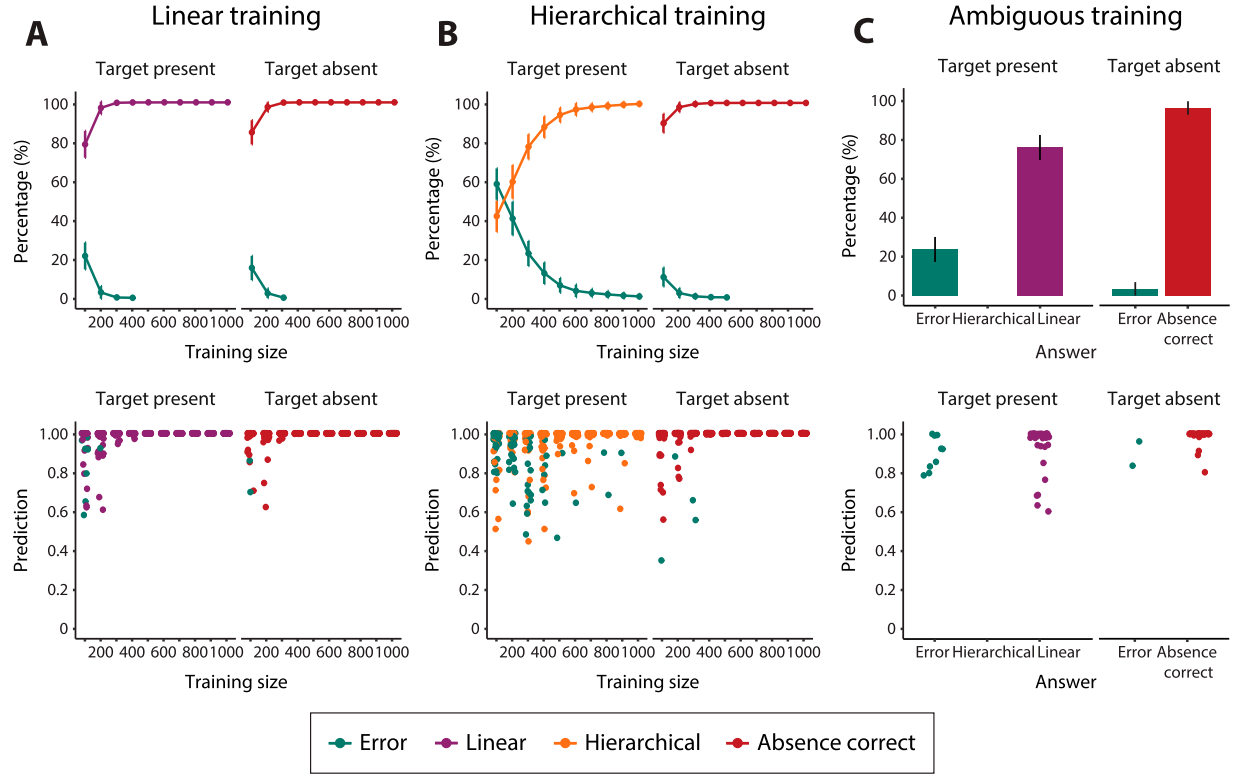

Figure 10. Model performance after linear training (a), hierarchical training (b) and ambiguous training (c). Results are into divided average accuracy over 100 simulations (error bars represent standard deviation) and specificity of predicted output (activation of output unit with largest value) on the test trials of one simulation.

2.3.2. Results

We evaluate each model's performance by comparing its predicted output on the test input to the correct test output. Each unit in the output layer of the model contains an activation value which can be interpreted as the likelihood that that unit corresponds to the position of the target, given the input (activations sum to one). We took the index of the output unit with the maximum activation value to be the model's predicted output. This value can be seen as the specificity of the model's prediction. For instance, if a model has learned to interpret the phrase *second blue ball* hierarchically, then given the picture in Figure 2(b) it outputs a vector with a high activation value for the fourth element (i.e. the target position, which has a one in the one-hot output vector used during training) and low activation values for all the other elements. We show the specificity of the predicted output (bottom graphs in Figure 10 each show these predictions for the test trials of one simulation), and evaluate the accuracy of these predictions by comparing them to the correct output (i.e. top graphs in Figure 10 show the average percentage correct: frequency with which the predictions match their labels).

When the model was trained on linear data, it quickly reached very good performance. After 400 training trials, the model scored perfectly, reaching an average accuracy of 100% (Figure 10(a)). After training sizes of 100 and 200, the model makes on average, respectively 19 and 3 errors. These all have to do with the presence of the target: the model either gives a target-absent response on a target-present trial (i.e. "miss"), or it gives an incorrectly linear response on a target-absent trial.

After 100 hierarchical training trials, the model reaches an average accuracy of 65%. The majority of its errors are wrong (but not linear) answers on target-present trials. The model's performance steadily increases with increasing training size up to 700 trials, after which it stabilises around 97–100% correct on target-present trials (Figure 10(b)). The hierarchical model needs more training data to reach high accuracy than the linear model, which has likely to do with the statistical variance in the hierarchical output data: whereas *second blue ball* on linear target-present trials always maps to position 2, the same target on hierarchical target-present trials can be in positions 3–8. More generally, the effect of hierarchy on interpretation in terms of statistics (i.e. in the form of an input-output mapping in our experiment) is inconsistent because it reflects information that is not directly encoded in the linear properties of the (input or output) signal.

In order to evaluate whether the model gives more linear or more hierarchical answers after being trained on ambiguous data, we simulated this evaluation 100 times. The model was trained on 100 different datasets of 100 ambiguous (convergent) trials, and at each simulation evaluated on 100 unambiguous (divergent) test trials. The model gets absence correct on most target-absent trials ($M = 96.5$, $SD = 3.47$), see Figure 10(c). Importantly, on target-present trials it gives mainly linear answers ($M = 76.2$, $SD = 6.43$), and never gives a hierarchical answer (see the empty column for "hierarchical" in Figure 10(c)). On average, the model makes 14 errors, which are of the same type as those made by the "linear" model (i.e. misses, or incorrectly linear answers on target-absent trials).

To evaluate how much unambiguously hierarchical information the model needs to start generalising hierarchically, we trained it on a mixed dataset with different ratios between ambiguous and unambiguously hierarchical trials. This ratio ranged from 10:0 (fully ambiguous) to 0:10 (fully hierarchical). Note that these mixed training data are always fully compatible with the hierarchical interpretation. What varies is the number of trials that is also compatible with the linear interpretation. Each mixed training set contained 100 trials, and we simulated each training-test evaluation 100 times. Figure 11 presents the responses for each of the different ratios. What is clear from the figure is that the more unambiguous evidence for the hierarchical interpretation in the training set, the more the model converges on the hierarchical interpretation in the test set. What is notable is that this increase is gradual: there is never a point at which the model "realises" that the hierarchical interpretation is the only correct generalisation (i.e. the model does not induce a rule). Instead, it always gives a substantial proportion of linear answers, even when 90% of the training data is unambiguously hierarchical and only 10% is ambiguous. Moreover, the number of errors on target-present trials increases as there is more unambiguous evidence for the hierarchical interpretation. This matches the patterns seen after linear and hierarchical training. The model initially only considers the linear interpretation, on which it does not make many errors (cf. Figure 10(a)), but the increasing evidence for the hierarchical interpretation is also taken as increasing evidence against the linear interpretation, so the model will give less linear responses. However, it still does not always get the hierarchical answer right, which is why its error rate increases (cf. Figure 10(b)). In all, these results again show that the model can learn to answer "hierarchically", but that it needs (a considerable percentage

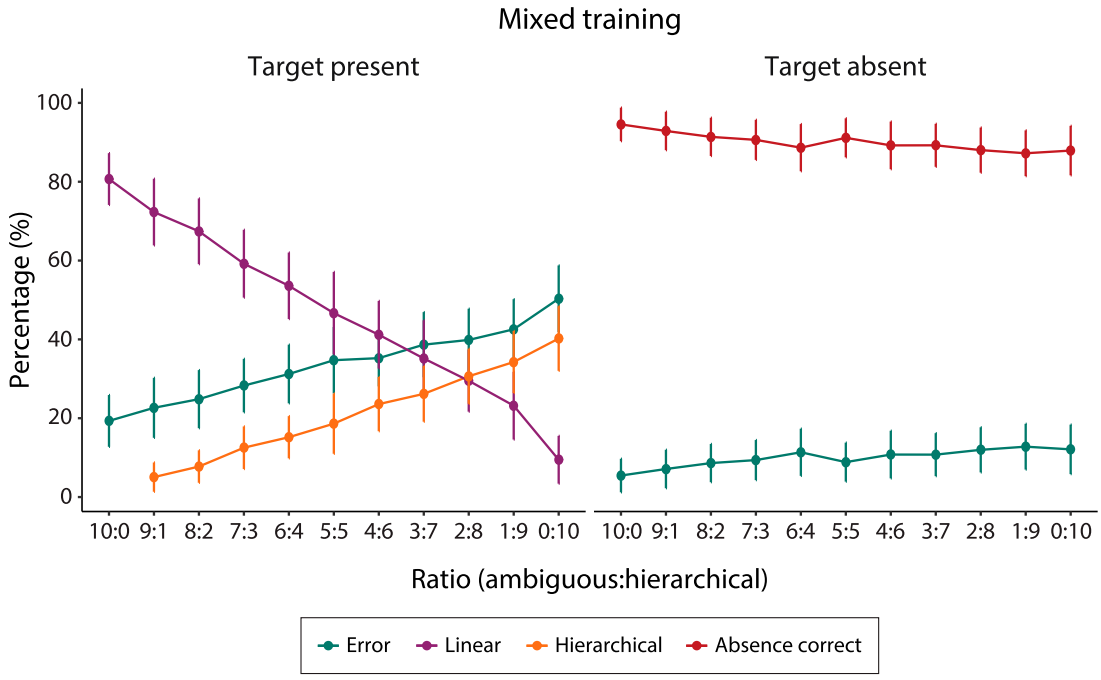

Figure 11. Model performance on hierarchical test trials after mixed training data. The training sets are composed of different ratios between ambiguous and unambiguously hierarchical trials, ranging from fully ambiguous data (10:0) to fully hierarchical data (0:10).

of) unambiguous trials to overcome its non-hierarchical bias.

2.3.2.1. Extrapolation and interpolation. We then probed the hierarchical model's ability to extrapolate and interpolate to novel items that were not seen during training. Figure 12 presents the model's accuracy, defined as the percentage of correct hierarchical answers, on both generalisation tests as a function of the input vectors that represented the words in the phrases. On the extrapolation test, the model did not generalise very well, regardless of whether it was trained on one-hot vector representations (Mean accuracy = 12.6, $SD = 9.20$), reduced word embeddings ($M = 12.7$, $SD = 5.44$) or full word embeddings ($M = 10.7$, $SD = 5.89$). In order to see whether these accuracies differ from chance level, we ran 100 simulations in which the training data consisted of pseudorandom mappings between input (phrase, picture) and output (target position). These contained the same information as the other simulations, and included one-hot vectors as the input layer. The model was tested on "third red ball". Given that there are 6 attested outputs in the hierarchical training regime for the ordinal "third" (i.e. positions 4 through 9), and that there is no consistent statistical relationship between the target and the output (i.e. there is nothing to learn, beyond the fact that "third" cannot be in the positions 1-3), this model scores around chance level of 16.7% accuracy. Comparison of the four groups (one-hot, reduced embeddings, full embeddings, random) through a one-way ANOVA in R (R Core Team, 2020) reveals that the accuracies between groups were different, $F(3,396) = 16.7$, $p < .001$, but pairwise follow-up tests showed that none of the conditions scored above chance. In fact, they all scored slightly below chance: one-hot vs. random: $\Delta = -4.63$, 95% CI [−7.12, −2.15], $p < .001$; full word embedding vs. random: $\Delta = -6.53$, 95% CI [−9.01, −4.05], $p < .001$; reduced word embedding vs. random: $\Delta = -4.53$, 95% CI [−7.01, −2.05], $p < .001$.

On the interpolation test, where the model is tested on "third red ball", it reached higher accuracy for each type of input vector: one-hot vectors ($M = 13.9$, $SD = 13.1$), reduced word embeddings ($M = 22.2$, $SD = 12.4$) and full word embeddings ($M = 23.8$, $SD = 22.2$). We again consider chance level to be around 16.7%, because the input "third red ball" during training could only be followed by a one-hot output vector with a one in either of the six positions 4–9. To evaluate each model against this chance level, we computed the model's performance after it was trained on pseudorandomly generated data, as described above. Comparison of the four groups (one-hot, reduced embeddings, full embeddings, random) again reveals that the accuracies between groups were different, $F(3,396) = 15.5$, $p < .001$. Pairwise follow-up tests showed that the accuracy for the full and reduced word embeddings was higher than expected by chance (full word embedding

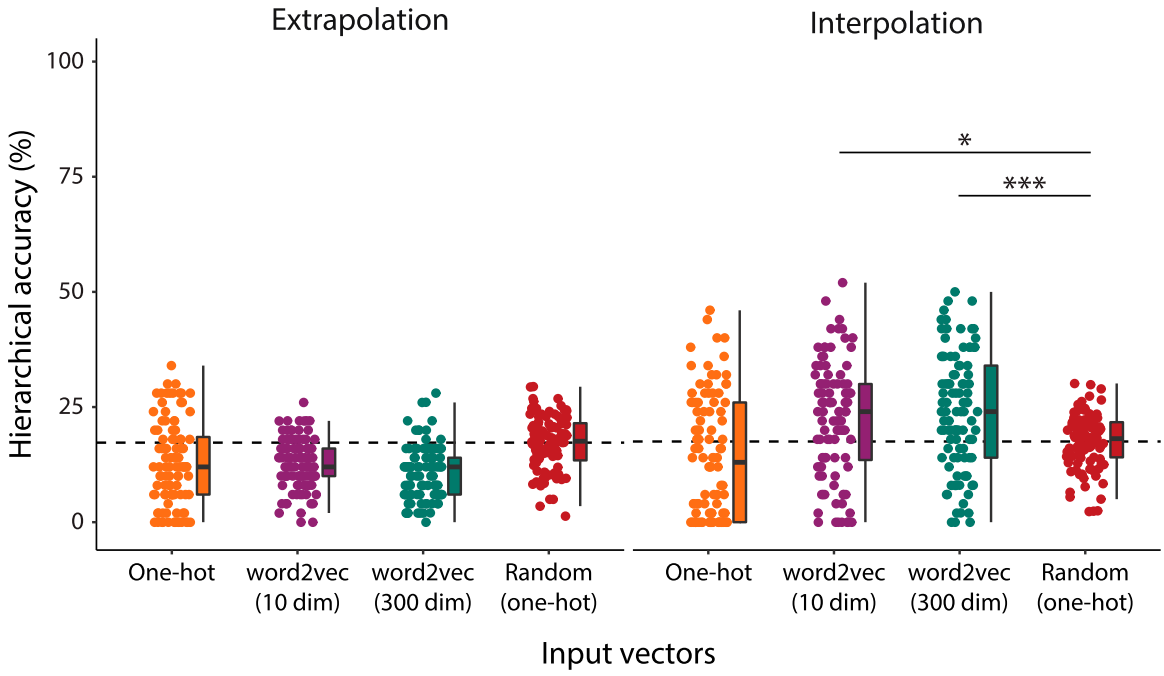

Figure 12. Percentage of correct hierarchical responses on both generalisation tests after training and testing on different input vectors. Each drop reflects the average hierarchical accuracy on one simulation run (100 simulations per evaluation). The horizontal line reflects the mean accuracy of the model after pseudorandom training, thus representing chance level.

vs. random: $\Delta = 6.32$, 95% CI [2.15, 10.5], $p < .001$; reduced word embedding vs. random: $\Delta = 4.66$, 95% CI [0.49, 8.84], $p = .02$). Interpolation accuracy for one-hot vectors was not different from chance. Despite this slight increase in accuracy for the model when trained on word embeddings, overall these findings show that the model was not able use the information it had induced from hierarchical training to systematically generalise to unseen items.

3. Discussion

In two behavioural experiments, we show a strong preference for hierarchy in human language interpretation: people's interpretation of ambiguous noun phrases categorically follows from their hierarchically organised, right-branching syntactic structure. In line with a long tradition of research, our findings support the idea that humans represent noun phrase structures in terms of hierarchical relations rather than linear order (Alexiadou et al., 2007; Cinque, 2005; Culbertson & Adger, 2014; Hamburger & Crain, 1984; Jackendoff, 1972; Martin et al., 2020; Pinker, 1999). In addition, we trained and tested an LSTM model on a computational version of the experimental task, and showed that the model can learn to give hierarchical answers if it is trained on unambiguously hierarchical datasets. However, when the training data contain both unambiguously hierarchical as well as ambiguous trials, the model strongly favours the linear interpretation, even though the hierarchical interpretation is a better fit to the overall data. Moreover, the "hierarchical" model does not systematically generalise to novel items that are not seen during training. These findings show that the model behaves unlike humans when the training data are ambiguous, and suggest that it needs different inductive biases in order to achieve human-like generalisation.

A comparison between the performance of the model and the behaviour of the human participants reveals a number of critical differences. First of all, while the model learned to give hierarchical answers, it only did so when it was explicitly fed unambiguously hierarchical information during supervised training. When the training data were ambiguous with respect to the correct representation underlying the noun phrases, the model had a strongly linear bias, never giving a hierarchical answer during the test phase. When the training data were mixed to contain both ambiguous and unambiguously hierarchical trials, such that the hierarchical interpretation was the only generalisation fully compatible with the data (i.e. the linear interpretation was only compatible with ambiguous trials), the model still had a strongly linear bias. This suggests that the model can learn to answer "hierarchically", but that it needs a substantial percentage of unambiguous trials to overcome its non-hierarchical bias (cf. McCoy et al., 2018).

The point about the apparent need for unambiguously hierarchical information during supervised training is relevant because children are not taught to interpret language hierarchically, but come to do so naturally, despite strongly deficient and ambiguous input data (e.g. Berwick et al., 2011; Crain, 1991; Gleitman & Newport, 1995; Kam & Fodor, 2012; Legate & Yang, 2002; Lidz et al., 2003). While we do not believe that adult language users have not been exposed to unambiguous data, it does seem to be the case that humans have a bias to interpret language in accordance with its underlying hierarchical structure (Crain & Nakayama, 1987; Crain et al., 2017; Ferrigno et al., 2020; Flaherty et al., 2021; Hunsicker & Goldin-Meadow, 2012; Kam & Fodor, 2012; Martin et al., 2020; Yang et al., 2017). The effect of such biases is particularly clear when people consistently generalise over hierarchical structure rather than linear order, despite the fact that these generalisations are underdetermined by the training data (Culbertson & Adger, 2014; Ferrigno et al., 2020; Martin et al., 2020; Morgan & Ferreira, 2021). This learnability scenario also applies to the interpretation of phrases such as *second blue ball*, even when the input does contain unambiguous data. There might indeed be positive evidence in the linguistic input to suggest that such a phrase should be interpreted as a hierarchical structure, but this does not yet rule out the interpretation derived from a linear structure. As is the case in most linguistic examples of ambiguity, evidence for interpretation A is not necessarily evidence against interpretation B. In our behavioural experiments, however, participants categorically interpreted *second blue ball* hierarchically, completely ignoring the linear interpretation, even though that linear option was always present. The strong preference to interpret these phrases hierarchically is suggestive of an inductive bias for hierarchy. Computational models without such an inductive hierarchical bias will often interpret ambiguous linguistic input in line with the linear generalisation, because that is the simpler statistical mapping between input and output sequence (Frank et al., 2013; McCoy et al., 2018, 2020). Indeed, it has been shown that RNNs have an architectural bias for dependencies over shorter (linear) distance (Christiansen & Chater, 1999).

In addition, the hierarchical model was not capable of systematic generalisation to novel items. We showed this by evaluating its ability to extrapolate (i.e. generalise to "third red ball" when the training data does not contain "red") and interpolate (generalise to "third red ball" when the training data contains "third", "red" and "ball", but not in combination) as a function of different types of input vectors (i.e. one-hot vectors and word embeddings). On the extrapolation test the model did not perform above chance level, even if it was trained and tested on word embeddings from word2vec (Mikolov et al., 2013). This is in line with previous studies which show that RNNs are not able to generalise to items that are not observed during training (Hupkes et al., 2020; Lake & Baroni, 2018; Loula et al., 2018), a consequence of the training algorithm also called "input independence" (Marcus, 1998, 2001). Note that the model's responses to items with the word "red", while labelled as errors, are technically not incorrect. Because the training data never contained "red" as possible input, every induction for a new item containing "red" is statistically legitimate (Marcus, 1998, 2001). Importantly, however, they differ sharply from what humans do. That is, modification in natural language is systematic, in that it applies in the same way to all variables of the right type. If someone knows how to interpret "second *blue* ball" and "second *green* ball", they interpret "second *red* ball" in a similar way, even if they have never seen "red" as a possible attribute. A well-known example of the productive and systematic nature of linguistic knowledge is children's behaviour on the Wug Test: young children know that the plural form of a pseudoword such as *wug* would be *wugs*, even though they have never heard this word before (Berko, 1958).

On the interpolation test, we found that if the model was trained on one-hot vectors, it performed at chance level. When it was trained on word embeddings, however, it scored somewhat higher than chance level, suggesting that it was able to take advantage of the inherent similarity between the word embeddings that represent related words, such as "blue" and "red". In addition, it is possible that the model picks up the statistical information that "red" and both "green" and "blue" occur in the same distributional environments, which would allow it to interpret "third red ball" correctly. However, we again believe that the reason behind this performance differs in a fundamental way from the reason why human cognition can support interpolation (and extrapolation). Human knowledge of linguistic modification relies on a symbolic representation of the way in which ordinals modify their arguments (i.e. ordinal(x); see the lambda expressions in (1)-(8)), which is why this relation obeys consistency and systematicity. The answers of the participants in our behavioural experiments were categorical: they consistently interpreted the phrases in the same way. The model's performance, instead, is stochastic: it gets the answer to "third red ball" right on about one-fourth of the trials, while making an error on all the other trials. The fact that the model was not able to consistently draw the right

generalisations (i.e. the highest average accuracy was 23.8% for the full word embeddings, but even this model sometimes reached 0% accuracy, see Figure 12) shows that the model was not capable of systematic generalisation. Rather, in line with previous work, it appears that the model is to some extent capable of generalising in an item-based manner, correctly interpreting novel items when they are composed of known features (Baroni, 2020; Lake & Baroni, 2018; Loula et al., 2018).

Importantly, the model's inability to systematically generalise to unseen items shows that it achieved its performance on hierarchical test trials without resorting to hierarchical constituent structure (Fodor & Pylyshyn, 1988; Marcus, 2001; Pinker, 1999; Pinker & Prince, 1988). To be clear, this is not to say that hierarchical structure per se is necessary for a system to be able to generalise. A computational system that relies only on linearly structured representations might be able to generalise, certainly if these representations contain symbolic variables to which specific instances can be bound. Our point is that the inability to systematically generalise to novel items suggests that the model does not rely on the type of symbolic constituent structure we believe underlies the responses of the human participants (Martin, 2020; Martin & Doumas, 2017, 2019; Puebla et al., 2021).

To sum up, we showed that an LSTM learns to provide output that is in line with hierarchical representations. However, the way in which the model generalises is quite different from linguistic generalisation by humans: when given ambiguous training data, it never provided hierarchical answers, and when tested on novel items, it did not systematically generalise. These two limitations show that the model's inductive biases and its ostensibly hierarchical knowledge are fundamentally different from human knowledge of language.

3.1. Linear models of hierarchical structure

While many contemporary computational models of language achieve impressive performance on a range of language tasks (e.g. machine translation, question answering), they often break down when evaluated on targeted syntactic tests. The reason is that they are fundamentally sequence-based models: they map one *sequence* onto another *sequence* (hence the term seq2seq models; Sutskever et al., 2014), and thus learn sequentially organised statistical patterns that cannot capture the full complexity of hierarchical syntax. While statistical signatures of hierarchical constituent structure can be found in the sequential structure of a sentence (e.g. Thompson & Newport, 2007), and while sequential statistics affects language processing (e.g. Townsend & Bever, 2001), that is not to say that sequence statistics is a sufficient basis for language (Chomsky, 1957). Because these models are inherently linear, they do not have a natural way to capture structural ambiguities (e.g. that *she saw the man with binoculars* has two meanings) and structural generalisations between different constructions (e.g. how *what did she see the man with?* relates to only one of these two meanings), which follow from the structured nature of linguistic representations and the structure dependence of linguistic operations.

In addition, the strongly linear bias of these computational models does not readily explain why structure dependence is so pervasive (Berwick et al., 2011; Crain & Pietroski, 2001; Fodor & Crowther, 2002; Heinz & Idsardi, 2011). If statistical information about sequential properties, such as linear order, were available as the basis for grammatical acquisition, one would expect speakers to adopt linear procedures, and therefore languages with linear dependencies to emerge, because that type of information is abundantly available. For instance, in the large majority of subject-verb agreement dependencies, the subject noun and the verb are adjacent. A language model which is trained and tested on these data can thus predict the correct verb inflection in most cases without accessing syntactic structure (Linzen et al., 2016). When the model is tested on structurally more complex examples, which are less likely to be found in the training data and which require hierarchical structure, its accuracy drops dramatically (Marvin & Linzen, 2018). Yet for humans this never happens: children universally adopt structure-dependent rules in the face of overwhelming evidence that is in line with linear alternatives (e.g. Crain & Nakayama, 1987; Crain et al., 2017; Gleitman & Newport, 1995; Lidz et al., 2003; Yang et al., 2017).

We noted in the previous sections that under the experimental circumstances in which the model was tested, it appears that statistical analysis of sequentially presented data is not sufficient to model human language behaviour (see also Puebla et al., 2021). This divergence between model performance and human behaviour could be attributed to roughly two independent factors: differences in cognitive architecture and differences in input data. Regarding input data, we acknowledge that the training data for computational models usually consists of raw texts, which lack rich sources of information that contribute to disambiguating the intended meanings of utterances (see Bender & Koller, 2019 for discussion). While this limits the generalisability of our findings in the same way as it limits most NLP work, we do recognise that other NLP models are trained on much more and more diverse data than what we used in our simulations. It is certainly possible that the LSTM would have performed

differently had it been trained on more naturalistic data. Assuming that naturalistic language data contains more evidence in favour of hierarchical structure, we predict that the model's performance on divergent test trials will reveal a stronger preference for the hierarchical interpretation, in line with what we show in our mixed training-test regime (see Figure 11).

That being said, even within the limited scope of our training simulation we showed that the model learned to behave "hierarchically". It was only after further investigation (in particular, extrapolation and interpolation) that we concluded that this behaviour did not arise in the same way as the linguistic behaviour of our participants. The difference in quality and quantity training, therefore, does not undermine our argument that hierarchical performance is not directly indicative of human-like hierarchical representations. We believe that progress towards human-like linguistic generalisation will benefit from a significant adjustment to the cognitive architecture of these models, such that they are biased to encode constituent structure (for related proposals, see Guest & Martin, 2021; Linzen, 2020; Linzen & Baroni, 2021). This might eventually turn out to be unnecessary in the sense that a preference for constituency could be learned from the environment, so it need not be innate (e.g. Perfors et al., 2011). Our current results do not speak to the question of innateness. However, what is crucial is not whether these biases are innate or learned, but whether they precede the acquisition of specific grammatical properties. Given the evidence for structure-dependent generalisations in both child and adult linguistic behaviour (Crain & Nakayama, 1987; Crain et al., 2017; Flaherty et al., 2021; Gleitman & Newport, 1995; Hunsicker & Goldin-Meadow, 2012; Kam & Fodor, 2012; Lidz et al., 2003; Martin et al., 2020; Yang et al., 2017), we believe that the incorporation of a notion of hierarchy into computational language models is the logical next step in order to build plausible models of human cognition.

In support of the value of this idea, recent results show that endowing neural networks with (syntactic) inductive biases for hierarchy improves their performance on complex syntactic tasks (e.g. Chen et al., 2017; Hale et al., 2018; Kuncoro et al., 2018; McCoy et al., 2020; Shen et al., 2019; Wilcox et al., 2019). These biases can be implemented in several ways, by means of both implicit and explicit representations of hierarchy. As an example of the former, the Ordered Neurons LSTM has an architecture in which its memory cells are structurally ordered in such a way that when a higher ordered neuron is updated, lower ordered neurons are forced to be updated as well. Different neurons therefore vary in update frequency, due to which they also vary in the timescale of the information they encode, with higher ordered neurons encoding longer timescales (Shen et al., 2019). As higher nodes in a tree structure represent information spanning over longer timescales, higher ordered neurons learn to encode higher nodes. This network thus comes to represent the hierarchical structure of sentences by discovering an implicit connection between timescale and node height. In contrast to this fully data-driven approach, the Tree-LSTM model is built to represent the hierarchical structure of sentences explicitly (Chen et al., 2017). This model is given the correct syntactic tree structure for every input sentence, such that its internal representations are biased to encode constituent structure. In contrast to the implicit link between node height and timescale in the Ordered Neurons LSTM (Shen et al., 2019), the Tree-LSTM incorporates syntactic trees explicitly (Chen et al., 2017). An important similarity between the two approaches, however, is that they both rely on the modeller's assumptions about the type of structure that must be represented.

3.2. Structure, statistics, or both?

A commonly articulated reason to favour linearity is that hierarchical structure is complex. Therefore, if language use can be equally well captured by a purely linear system, the linear system should be favoured on grounds of parsimony (e.g. Frank et al., 2012; Frank & Christiansen, 2018). However, while the hierarchical structure of natural language syntax is indeed more complex than can be modelled by linear grammars (Chomsky, 1956), equivalent metrics of parsing complexity have not been defined for "linear" vs. "hierarchical" language use. Thus, without an implementation of syntactic structure building, or at least the identification of the core computations at stake, the simplicity statement is ill-posed. Furthermore, the psycholinguistic evidence that hierarchical structure building is costly comes from the comparison of putatively "more complex" structure with "less complex", but still hierarchical, structure (e.g. King & Just, 1991; Waters & Caplan, 2004). But most importantly, appeals to simplicity can only be made when competing theories have equivalent empirical coverage, which does not hold here because the "linearity view" cannot account for our behavioural results. To deal with these situations, hierarchical structure might only be used in very specific situations, such as when sentence meaning depends on precise hierarchical structure (*second blue ball* vs. *big blue ball*; e.g. Frank et al., 2012). However, this requires the postulation of both a linear and a hierarchical grammar processor, resulting in a two-system cognitive architecture that is

more complex than a one-system architecture that only uses hierarchical syntax (Lewis & Phillips, 2015).

While the debate about hierarchical and linear systems is often couched in terms of hierarchy versus statistics, these two are not mutually exclusive (see Martin, 2016, 2020; Yang, 2004). We believe that probabilistic processes do play an important role in language, as has been shown extensively (e.g. Marcus, 2001; Pinker, 1999; Townsend & Bever, 2001), but that they operate within the boundaries imposed by hierarchical structure, during both language processing (Martin, 2016, 2020) and acquisition (Lidz & Gagliardi, 2015; Yang, 2002, 2004). Finding out where the boundaries lie, i.e. what is the representational level over which probabilities are computed, is an important avenue for future research (Brennan et al., 2016; Brennan & Martin, 2020; Martin & Doumas, 2017, 2019; Meyer et al., 2019).

4. Conclusion

In conclusion, we have shown that hierarchical structure is a key component of human language interpretation, and that an LSTM only reproduces such hierarchical behaviour under highly specific training circumstances. We conclude that without a predisposition to generalise hierarchically, the model is not a cognitively adequate model of human language (Fitch, 2014; Martin, 2020; Martin & Doumas, 2017, 2019, 2020). Beyond language, hierarchical structure might form the basis of other domains of cognition and information processing (e.g. Dehaene et al., 2015; Doumas et al., 2008; Ferrigno et al., 2020; Fitch, 2014; Hummel & Holyoak, 1997; Martin & Doumas, 2019, 2020; Tenenbaum et al., 2011). Figuring out how the brain builds hierarchically structured representations from linear input therefore remains a central question in the science of the human mind.

Notes

1. Semantic scope refers to the domain in which an operator can affect the interpretation of other elements. Scope domains can sometimes be directly read off hierarchical relations between syntactic elements, that is, by virtue of the c-command relation (e.g., Reinhart, 1983).
2. Notably, even if it is the case that such pragmatic factors drive people to interpret *second blue ball* non-intersectively (i.e., not referring to the ball that is blue and in second position), they would still have to use constituent structure to interpret the phrase "hierarchically".
3. This applies to all ordinals except the ordinal *first*, for which divergent trials do not exist. That is, if the first among blue balls is not the first ball, then the linear interpretation is not present and only the hierarchical option is available. Divergent trials with ordinal *first* were actually non-convergent trials with only a hierarchical option, and were therefore not analysed.
4. Again, the interpretations of targets with the ordinal *first* always converged. The responses to these targets could not distinguish between the two interpretations and were therefore not analysed.
5. Code and data for the computational experiments described in this paper are available at: https://github.com/CasCoopmans/second_blue_ball.
6. Trials with the ordinal *first* would always be convergent and were therefore not part of the datasets. We replaced *first* by *seventh* in these datasets to make sure that the number of ordinals on which the human participants and the model were tested was the same.
7. This was done to make sure that the number of target-present and target-absent responses for each ordinal are roughly equal. However, it also means that output vectors with a one in position 9 ("target absent") are overrepresented in the output. We accounted for this imbalance by updating the loss function with a weighting parameter that reflected the class distribution in the training data (Chollet et al., 2015).
8. Because we have ten words, we would need maximally ten dimensions to capture their differences. In reality, this number can be lower, because some of the words are related.

Acknowledgements

We thank Stefan Frank and Hartmut Fitz for their comments on an earlier version of this manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Andrea E. Martin was supported by the Lisa Meitner Research Group "Language and Computation in Neural Systems" from the Max Planck Society and by the Netherlands Organisation for Scientific Research (NWO) [grant number: 016.Vidi.188.029]. Peter Hagoort was supported by the NWO Gravitation Grant [grant number: 024.001.006] to the Language in Interaction Consortium.

ORCID

Cas W. Coopmans http://orcid.org/0000-0001-7622-3161
Helen de Hoop http://orcid.org/0000-0001-8652-1146
Peter Hagoort http://orcid.org/0000-0001-7280-7549
Andrea E. Martin http://orcid.org/0000-0002-3395-7234

References

Alexiadou, A., Haegeman, L., & Stavrou, M. (2007). *Noun phrase in the generative perspective*. Mouton de Gruyter.

Baroni, M. (2020). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375* (1791), 20190307. https://doi.org/10.1098/rstb.2019.0307

Bender, E. M., & Koller, A. (2019). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198).

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*(2–3), 150–177. https://doi.org/10.1080/00437956.1958.11659661

Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science*, *35*(7), 1207–1242. https://doi.org/10.1111/j.1551-6709.2011.01189.x

Brennan, J. R., & Martin, A. E. (2020). Phase synchronization varies systematically with linguistic structure composition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190305. https://doi.org/10.1098/rstb.2019.0305

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157–158*, 81–94. https://doi.org/10.1016/j.bandl.2016.04.008

Bybee, J. (2002). Sequentiality as the basis of constituent structure. In T. Givón & B. F. Malle (Eds.), *The evolution of language out of pre-language* (pp. 109–134). John Benjamins.

Chen, H., Huang, S., Chiang, D., & Chen, J. (2017). Improved neural machine translation with a syntax-aware encoder and decoder. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Long Papers* (pp. 1936–1945).

Chollet, F., et al. (2015). Keras. https://keras.io

Chomsky, N. (1956). Three models for the description of language. *IEEE Transactions on Information Theory*, *2*(3), 113–124. https://doi.org/10.1109/TIT.1956.1056813

Chomsky, N. (1957). *Syntactic structures*. Mouton.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*(3), 157–205. https://doi.org/10.1207/s15516709cog2302_2

Christiansen, M. H., & Chater, N. (2015). The language faculty that wasn't: A usage-based account of natural language recursion. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01182

Christiansen, M. H., & MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning*, *59*(s1), 126–161. https://doi.org/10.1111/j.1467-9922.2009.00538.x

Cinque, G. (2005). Deriving Greenberg's Universal 20 and its exceptions. *Linguistic Inquiry*, *36*(3), 315–332. https://doi.org/10.1162/0024389054396917

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, *14*(4), 597–612. https://doi.org/10.1017/S0140525X00071491

Crain, S., Koring, L., & Thornton, R. (2017). Language acquisition from a biolinguistic perspective. *Neuroscience & Biobehavioral Reviews*, *81*, 120–149. https://doi.org/10.1016/j.neubiorev.2016.09.004

Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, *63*(3), 522–543. https://doi.org/10.2307/415004

Crain, S., & Pietroski, P. (2001). Nature, nurture and Universal Grammar. *Linguistics and Philosophy*, *24*(2), 139–186. https://doi.org/10.1023/A:1005694100138

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, *111*(16), 5842–5847. https://doi.org/10.1073/pnas.1320525111

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The neural representation of sequences: From transition probabilities to algebraic patterns and linguistic trees. *Neuron*, *88*(1), 2–19. https://doi.org/10.1016/j.neuron.2015.09.019

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *Preprint: arXiv:1810.04805*.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1–43. https://doi.org/10.1037/0033-295X.115.1.1

Everaert, M. B. H., Huybregts, M. A. C., Chomsky, N., Berwick, R. C., & Bolhuis, J. J. (2015). Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, *19*(12), 729–743. https://doi.org/10.1016/j.tics.2015.09.008

Ferrigno, S., Cheyette, S. J., Piantadosi, S. T., & Cantlon, J. F. (2020). Recursive sequence generation in monkeys, children. U.S. Adults, and native Amazonians. *Science Advances*, *6*(26), eaaz1002. https://doi.org/10.1126/sciadv.aaz1002

Fitch, W. T. (2014). Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition. *Physics of Life Reviews*, *11*(3), 329–364. https://doi.org/10.1016/j.plrev.2014.04.005

Flaherty, M., Hunsicker, D., & Goldin-Meadow, S. (2021). Structural biases that children bring to language learning: A cross-cultural look at gestural input to homesign. *Cognition*, *211*, 104608. https://doi.org/10.1016/j.cognition.2021.104608

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Fodor, J. D., & Crowther, C. (2002). Understanding stimulus poverty arguments. *The Linguistic Review*, *18*(1–2), 1–2. https://doi.org/10.1515/tlir.19.1-2.105

Frank, R., Mathis, D., & Badecker, W. (2013). The acquisition of anaphora by simple recurrent networks. *Language Acquisition*, *20*(3), 181–227. https://doi.org/10.1080/10489223.2013.796950

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*(6), 829–834. https://doi.org/10.1177/0956797611409589

Frank, S. L., Bod, R., & Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, *279*(1747), 4522–4531. https://doi.org/10.1098/rspb.2012.1741

Frank, S. L., & Christiansen, M. H. (2018). Hierarchical and sequential processing of language. *Language, Cognition and Neuroscience*, *33*(9), 1213–1218. https://doi.org/10.1080/23273798.2018.1424347

Gleitman, L., & Newport, E. (1995). The invention of language by children: Environmental and biological influences on the acquisition of language. In L. R. Gletiman & M. Lieberman (Eds.), *An invitation to cognitive science*, *Language* (Vol. 1, pp. 1–24). MIT Press.

Goldberg, A. E., & Michaelis, L. A. (2017). One among many: Anaphoric one and its relationship with numeral one. *Cognitive Science*, *41*(S2), 233–258. https://doi.org/10.1111/cogs.12339

Goldin-Meadow, S. (2003). *The resilience of language: What gesture creation in deaf children can tell us about how all children learn language*. Psychology Press.

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (Ed.), *Universals of language* (pp. 73–113). MIT Press.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colourless green recurrent networks dream hierarchically. In *Proceeding of the 2018 North American Chapter of the Association for Computational Linguistics* (pp. 1195–1205).

Hale, J. T., Dyer, D., Kuncoro, A., & Brennan, J. (2018). Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Long Papers* (pp. 2727–2736).

Hamburger, H., & Crain, S. (1984). Acquisition of cognitive compiling. *Cognition*, *17*(2), 85–136. https://doi.org/10.1016/0010-0277(84)90015-5

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar*. Blackwell.

Heinz, J., & Idsardi, W. (2011). Sentence and word complexity. *Science*, *333*(6040), 295–297. https://doi.org/10.1126/science.1210358

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427–466. https://doi.org/10.1037/0033-295X.104.3.427

Hunsicker, D., & Goldin-Meadow, S. (2012). Hierarchical structure in a self-created communication system: Building nominal constituents in homesign. *Language*, *88*(4), 732–763. https://doi.org/10.1353/lan.2012.0092

Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2020). Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, *67*, 757–795. https://doi.org/10.1613/jair.1.11674

Jackendoff, R. (1972). *Semantic interpretation in generative grammar*. MIT Press.

Kam, X.-N. C., & Fodor, J. D. (2012). Children's acquisition of syntax: Simple models are too simple. In M. Piattelli-Palmarini & R. C. Berwick (Eds.), *Rich Languages From Poor inputs* (pp. 43–60). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199590339.003.0003

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*(5), 580–602. https://doi.org/10.1016/0749-596X(91)90027-H

Kuncoro, A., Dyer, D., Hale, J., Yogatama, D., Clark, S., & Blunsom, P. (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*, 1, Short Papers (pp. 1426–1436).

Lake, B., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2879–2888).

Legate, J. A., & Yang, C. D. (2002). Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review*, *18*(1–2), 1–2. https://doi.org/10.1515/tlir.19.1-2.151

Lewis, S., & Phillips, C. (2015). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research*, *44*(1), 27–46. https://doi.org/10.1007/s10936-014-9329-z

Lidz, J., & Gagliardi, A. (2015). How nature meets nurture: Universal Grammar and statistical learning. *Annual Review of Linguistics*, *1*(1), 333–353. https://doi.org/10.1146/annurev-linguist-030514-125236

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition*, *89*(3), 295–303. https://doi.org/10.1016/S0010-0277(03)00116-1

Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5210–5217).

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*(1), 195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535. https://doi.org/10.1162/tacl_a_00115

Loula, J., Baroni, M., & Lake, B. M. (2018). Rearranging the familiar: Testing compositional generalization in recurrent networks. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 108–114).

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, *37*(3), 243–282. https://doi.org/10.1006/cogp.1998.0694

Marcus, G. F. (2001). *The algebraic mind*. MIT Press.

Martin, A., Holtz, A., Abels, K., Adger, D., & Culbertson, J. (2020). Experimental evidence for the influence of structure and meaning on linear order in the noun phrase. *Glossa: A Journal of General Linguistics*, *5*(1), 97. https://doi.org/10.5334/gjgl.1085

Martin, A. E. (2016). Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in Psychology*, *7*, 120. https://doi.org/10.3389/fpsyg.2016.00120

Martin, A. E. (2020). A compositional neural architecture for language. *Journal of Cognitive Neuroscience*, *32*(8), 1407–1427. https://doi.org/10.1162/jocn_a_01552

Martin, A. E., & Doumas, L. A. A. (2017). A mechanism for the cortical computation of hierarchical linguistic structure. *PLOS Biology*, *15*(3), e2000663. https://doi.org/10.1371/journal.pbio.2000663

Martin, A. E., & Doumas, L. A. A. (2019). Predicate learning in neural systems: Using oscillations to discover latent structure. *Current Opinion in Behavioral Sciences*, *29*, 77–83. https://doi.org/10.1016/j.cobeha.2019.04.008

Martin, A. E., & Doumas, L. A. A. (2020). Tensors and compositionality in neural systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190306. https://doi.org/10.1098/rstb.2019.0306

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1192–1202).

Matthei, E. H. (1982). The acquisition of prenominal modifier sequences. *Cognition*, *11*(3), 301–332. https://doi.org/10.1016/0010-0277(82)90018-X

McCoy, R. T., Frank, R., & Linzen, T. (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* (pp. 2093–2098).

McCoy, R. T., Frank, R., & Linzen, T. (2020). Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, *8*, 125–140. https://doi.org/10.1162/tacl_a_00304

Meyer, L., Sun, Y., & Martin, A. E. (2019). Synchronous, but not entrained: Exogenous and endogenous cortical rhythms of speech and language processing. *Language, Cognition and Neuroscience*, *35*(9), 1089–1099. https://doi.org/10.1080/23273798.2019.1693050

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Preprint: arXiv:1301.3781*.

Morgan, A. M., & Ferreira, V. S. (2021). Beyond input: Language learners produce novel relative clause types without exposure. *Journal of Cognitive Psychology*, *33*(5), 483–517. https://doi.org/10.1080/20445911.2021.1928678

Mulligan, K., Frank, R., & Linzen, T. (2021). Structure here, bias there: Hierarchical generalization by jointly learning syntactic transformations. *Proceedings of the Society for Computation in Linguistics*, *4*(13), 125–135. https://doi.org/10.7275/j0es-xf97

Partee, B. (1975). Montague grammar and transformational grammar. *Linguistic Inquiry*, *6*(2), 203–300.

Partee, B. (2007). Compositionality and coercion in semantics: The dynamics of adjective meaning. In G. Bouma, I. Krämer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 145–161). Royal Netherlands Academcy of Arts and Sciences.

Payne, J., Pullum, G. K., Scholz, B. C., & Berlage, E. (2013). Anaphoric one and its implications. *Language*, *89*(4), 794–829. https://doi.org/10.1353/lan.2013.0071

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338. https://doi.org/10.1016/j.cognition.2010.11.001

Pinker, S. (1999). *Words and rules*. HarperCollins.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1), 73–193. https://doi.org/10.1016/0010-0277(88)90032-7

Puebla, G., Martin, A. E., & Doumas, L. A. A. (2021). The relational processing limits of classic and contemporary neural network models of language processing. *Language, Cognition and Neuroscience*, *36*(2), 240–254. https://doi.org/10.1080/23273798.2020.1821906

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org

Reinhart, T. (1983). *Anaphora and semantic interpretation*. Croom Helm.

Scheepers, C., & Sturt, P. (2014). Bidirectional syntactic priming across cognitive domains: From arithmetic to language and back. *Quarterly Journal of Experimental Psychology*, *67*(8), 1643–1654. https://doi.org/10.1080/17470218.2013.873815

Shen, Y., Tan, S., Sordoni, A., & Courville, A. (2019). Ordered neurons: Integrating tree structures into recurrent neural networks. In *Proceedings of the 7th International Conference on Learning Representations*.

Shlens, J. (2014). A tutorial on principal component analysis. *Preprint: arXiv:1404.1100*.

Spenader, J., & Blutner, R. (2007). Compositionality and systematicity. In G. Bouma, I. Krämer, & J. Zwarts (Eds.), *Cognitive Foundations of Interpretation* (pp. 163–174). Royal Netherlands Academcy of Arts and Sciences.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 3104–3112

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*(1), 1–42. https://doi.org/10.1080/15475440709336999

Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules*. MIT Press.

Tran, K., Bisazza, A., & Monz, C. (2018). The importance of being recurrent for modeling hierarchical structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4731–4736).

Waters, G. S., & Caplan, D. (2004). Verbal working memory and on-line syntactic processing: Evidence from self-paced listening. *The Quarterly Journal of Experimental Psychology Section A*, *57*(1), 129–163. https://doi.org/10.1080/02724980343000170

Wilcox, E., Qian, P., Futrell, R., Ballesteros, M., & Levy, R. (2019). Structural supervision improves learning of non-local grammatical dependencies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3302–3312).

Yang, C. D. (2002). *Knowledge and learning in natural language*. Oxford University Press.

Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, *8*(10), 451–456. https://doi.org/10.1016/j.tics.2004.08.006

Yang, C. D., Crain, S., Berwick, R. C., Chomsky, N., & Bolhuis, J. J. (2017). The growth of language: Universal Grammar, experience, and principles of computation. *Neuroscience and Biobehavioral Reviews*, *81*, 103–119. https://doi.org/10.1016/j.neubiorev.2016.12.023