Kallini et al. (2024) do not compare impossible languages with constituency-based ones

Tim Hunter
timhunter@ucla.edu

October 15, 2024

A central goal of linguistic theory, since at least Chomsky (1965, p.25), has been to find a precise characterization of the notion "possible human language". Researchers have pursued this goal by attempting to identify a kind of computational device that is capable of describing all *and only* the possible human languages, i.e. those languages that can be acquired by a typically developing human child. To the extent that a particular kind of computational device meets this goal, it constitutes a plausible hypothesis about the mental machinery that underlies the human capacity for language.

The success of recent "large language models" (LLMs) in NLP applications raises the possibility that LLMs might be devices that meet this goal. They have been found to be remarkably successful at tasks that, let us grant — controversially, but innocuously for present purposes — require learning certain human languages in a relevant sense. The other side of the coin, however, is whether LLMs are similarly successful at learning languages that humans cannot, i.e. "humanly impossible languages". If they are, this would tell against the hypothesis that human linguistic capacities take a form that resembles an LLM.

Kallini et al. (2024) cite a number of claims to the effect that LLMs will successfully learn such impossible languages, and set out to test this. They develop a set of synthetic languages with properties that are unlike what has been observed in any human language, and find that "GPT-2 struggles to learn impossible languages when compared to English as a control, challenging the core claim" (p.14691). The most interesting impossible languages, and the ones that Kallini et al. address most extensively in their paper, are those that involve count-based rules. Sentences of the language called WordHop, for example, are like sentences of English except that inflectional affixes on verbs are replaced with distinguished marker tokens (S for singular, P for plural) which appear to the right of the (uninflected) verb, separated by exactly four words; see (1). For a minimal comparison with WordHop, Kallini et al. also construct a minor variant of English called NoHop, which uses the same distinguished markers but places them immediately adjacent to the verb.

(1)

	Singular agreement example	Plural agreement example
English	He cleans his very messy bookshelf .	They clean his very messy bookshelf .
WordHop	He clean his very messy bookshelf S .	They clean his very messy bookshelf P .
NoHop	He clean S his very messy bookshelf .	They clean P his very messy bookshelf .

It is widely agreed that the count-based placement of the S and P markers in WordHop is indeed outside the bounds of "possible human languages" (whereas NoHop, being essentially analogous to English, is not), and Kallini et al.'s results show that GPT-2 is less successful at learning WordHop than NoHop. This finding is presented as the main challenge to the claims that GPT-2 models are insufficiently human-like.

The comparison between WordHop and NoHop, however, does not actually test the critical point. The problem, to a first approximation, is a confound between whether a rule is count-based and whether that rule creates non-adjacent dependencies: the comparison is between *adjacency* and *count-based non-adjacency*.

The crucial observation that linguists have repeatedly remarked on regarding count-based non-adjacent dependencies is their absence relative to *constituency-based non-adjacent* dependencies, not relative to adjacent dependencies. The corresponding claim about the human language faculty is that it can naturally accommodate or express constituency-based non-adjacent dependencies to a degree that does not hold for count-based non-adjacent dependencies. It would be interesting to know whether LLMs show this same asymmetry, but a comparison between WordHop and NoHop sheds no light on this question.

In section 1 I will rehearse some standard arguments illustrating the difference between count-based and constituency-based rules. With some specifics of the relevant phenomena in hand, section 2 lays out more carefully why the comparison between WordHop and NoHop misses the mark. This logic will lead to some suggestions for more appropriate comparisons in section 3.

1 Review of the underlying issues

The frequently-used example of question-formation in English provides a relevant entry point for illustrating the issues.[1] Consider the relationship that the sentences in (2a) and (3a) stand in to their corresponding yes-no questions. The question form of (2a) consists of the same words rearranged, as in (2b); we can describe the rearrangement by saying that the word 'will' has been displaced to the front of the sentence. One could imagine that this was an instance of a count-based rule that formed questions by displacing the third word of a sentence, but we can see that this is not the case because applying this count-based rule to (3a) yields (3b). The actual rule under investigation somehow yields (3c), where it is the sixth word that is displaced.

(2) a. The dog will bark
b. Will the dog bark?

(3) a. The dog in the corner will bark
b. * In the dog the corner will bark?
c. Will the dog in the corner bark?

Considering now (4a), the question-forming rule displaces neither the third word nor the sixth word (which would yield (4b) and (4c) respectively). What (2b) and (3c) have in common is that in both cases the displaced word is 'will', and this also holds of the desired form (4d) — where the displaced 'will' was the eighth word. But the rule under investigation somehow excludes moving the other 'will', the fourth word of (4a), to produce (4e).

(4) a. The dog that will chase the cat will bark
b. * That the dog will chase the cat will bark?
c. * The the dog that will chase cat will bark?
d. Will the dog that will chase the cat bark?
e. * Will the dog that chase the cat will bark?

And it is not as simple as always moving the last/rightmost occurrence of 'will' (or more generally, an auxiliary verb), as illustrated by the pattern in (5).

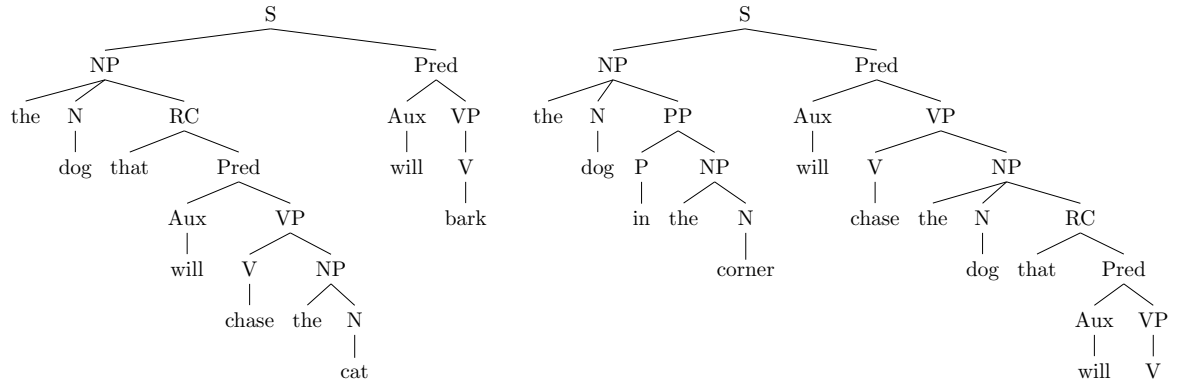
(5) a. The dog in the corner will chase the dog that will bark
b. Will the dog in the corner chase the dog that will bark?
c. * Will the dog in the corner will chase the dog that bark?

[1] This argument has appeared in numerous places, virtually unchanged, going back to at least Chomsky (1971, pp.26–29). Freidin (1991) gives a version that particularly emphasizes the contrast with count-based rules. Other sources include Chomsky (1975, pp.30–33), Chomsky (1980, pp.39–40) and Chomsky (1988, pp.41–45). For textbook expositions, see e.g. Akmajian et al. (2001, pp.156–168), Lasnik et al. (2000, pp.5–7) and Radford (1988, pp.31–34). Many of these sources discuss this question-formation rule as part of a "poverty of stimulus" argument, which need not concern us here: what's relevant here is just the initial point that *linguists* can test and disprove hypothesized count-based rules, not the subsequent question of how or why language-learners converge on the non-count-based rules that they do.

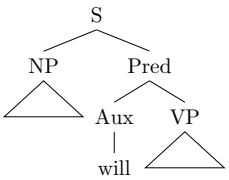
The operative rule cannot be formulated in count-based terms, i.e. no description of the form "the nth word of the sentence" or "the nth occurrence of 'will' from the end of the sentence" will consistently pick out the word that is to be displaced.

The correct generalization *can* be expressed in terms of hierarchical constituency: given the structural analyses in (6) for the declaratives in (4) and (5), the displaced word is the Aux that is the granddaughter of the root S node in the template in (7).

(6)



(7)



This example from English is entirely representative: patterns like this that conform to a constituency-based rule, but where no count-based characterization has been found, are ubiquitous in natural languages. And the reverse situation, where a pattern follows a count-based rule but has no constituency-based characterization, is unheard of. The conventional linguistic explanation for this striking asymmetry is that (languages with) count-based rules are "humanly impossible" — outside the capacity of the mental faculties that are recruited in naturalistic language development.[2] Of course, given a simple enough artificial grammar-learning experiment, a human may well show some success at learning and applying a count-based rule, perhaps by recruiting *other* mental faculties to the task; somewhat similarly, a proponent of the idea that LLMs embody a human-like ill-suitedness to count-based rules is not committed to the prediction that an LLM will always show precisely zero evidence of having extracted any count-based rule from training data. Rather than any raw measure of successful learning of any single kind of rule, the critical issue is an asymmetry between count-based and constituency-based rules.

Testing for such an asymmetry obviously requires controlling for other factors. While the rule for the placement of the S and P markers in Kallini et al.'s WordHop is a canonical example of a count-based rule — the kind that turns out to be insufficient to describe the pattern in (2)–(5) — the rule for placing these markers in NoHop is not an appropriately representative constituency-based rule to compare it against. The NoHop rule is extremely simple: the marker should be placed immediately after the verb. It's true that the full-fledged English system of verbal inflections involves crucially constituency-based rules, which are in fact closely intertwined with the classic phenomena in (2)–(5) above, and one of the configurations that this system produces is the one illustrated in (1), with the inflected verb 'cleans'. But the constituency-based parts of that system are not probed by a comparison between WordHop and NoHop, which differ *only* in whether the S and P markers are separated from the verb by four words or zero words.

[2]The idea is not that a count-based language would "die out" because of a failure on the part of human learners to perpetuate it; rather, the idea is that no human's linguistic development would ever give rise to such a language in the first place.

To flesh out this point, section 2 illustrates some of the constituency-based rules governing English verbal inflections that turn out to be independent of the differences between WordHop and NoHop. This will then lead to a proposal for a more appropriate comparison in section 3.

2 Constituency and English verbal inflections

A mistaken impression that NoHop can serve as a representative of constituency-based rules might arise, in part, from the fact that the behaviour of verbal inflections is intertwined with the question-forming rule that is used in the classical illustration of constituency-sensitivity rehearsed in section 1.

This connection can be established by observing that these inflections (e.g. the suffixes in 'cleans' and 'cleaned') do not co-occur with words like 'will' that are displaced by the question-forming rule. A finite clause must include either one of these inflections or a word that behaves like 'will' (e.g. 'may', 'must', 'can'), but not both.

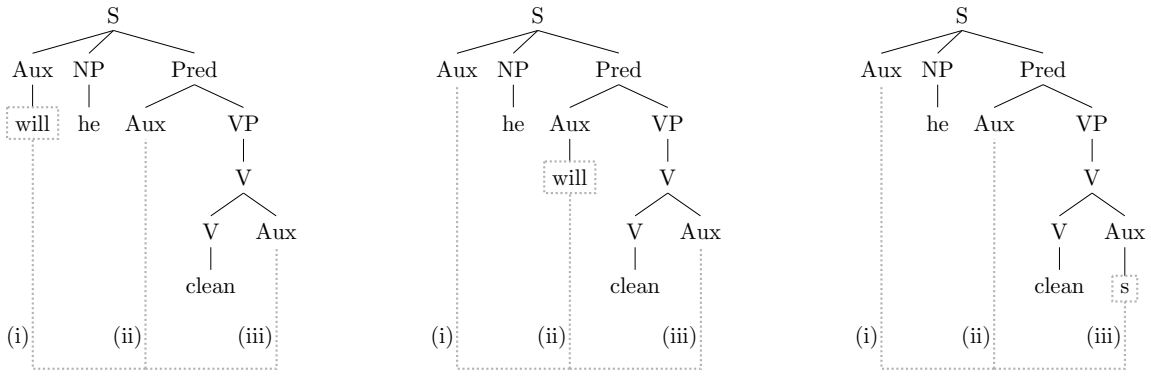
(8) a. * He clean
b. He will clean
c. He may clean

(9) a. He cleans
b. * He will cleans
c. * He may cleans

(10) a. He cleaned
b. * He will cleaned
c. * He may cleaned

So we have effectively identified a three-way dependency between (i) the sentence-initial position occupied by 'will' in the questions in section 1, (ii) the position occupied by 'will' in non-questions, in section 1 and in (8)–(10), and (iii) the position occupied by the inflectional affixes in (8)–(10). This can be formalized in various ways (see Chomsky (1957) for the original analysis along these lines[3]), but the dotted lines in (11) convey the key idea in a way that will suffice for our purposes here.

(11)



To complete the picture, notice that the question-forming rule does not make any distinction between the affixes that appear in position (iii) in declaratives and the words like 'will' that appear in position (ii): the affixes are also displaced to position (i) in questions, where their pronunciation is supported by a form of the dummy verb 'do'.

(12) a. Does he clean? (cf. (9a))
b. Did he clean? (cf. (10a))
c. Will he clean? (cf. (8b))
d. May he clean? (cf. (8c))

No matter how these details are formalized, the crucial and uncontroversial point is that these three interdependent positions are identified in constituent-based terms, not count-based terms. We saw in section 1 that the relationship between positions (i) and (ii) is not defined via a number of intervening words, but rather with reference to the hierarchical structure. Similarly, although the word that an affix in position (iii) attaches to has been adjacent to position (ii) in all the examples so far, this is not true in general: additional

[3]For textbook expositions see e.g. Fromkin et al. (2000, pp.259–300), Carnie (2007, pp.246–271), Lasnik et al. (2000, pp.66–86), Freidin (1992, pp.144-164).

words can intervene here too, as illustrated by (13). The presence of a direct object after the verb in these sentences also demonstrates that position (iii) can not be defined linearly as "the end of the string".

(13) a. He will without doubt clean his very messy bookshelf.
b. He without doubt cleans his very messy bookshelf.

Furthermore, although the discussion in section 1 emphasized only the hierarchical determination of the auxiliary that should be displaced to the front of the sentence, this target position is in fact defined in hierarchical terms too: (14) shows examples of questions where 'will' is in position (i) despite not being sentence-initial.[4]

(14) a. Which very messy bookshelf will he clean?
b. How will he clean his very messy bookshelf?
c. Though his bookshelf is very messy, will he clean it?

Another way in which English verbal inflections are intertwined with crucially hierarchical notions concerns number agreement with the subject; recall the two columns in (1). There is a single hierarchically-defined position that the agreement-controlling noun 'gift(s)' occupies in all of the examples in (15)–(16). The rule needs to pick out the second word (and the first of the two nouns) in (15), but the third word (and the second of the two nouns) in (16), so again no count-based formulation is possible.

(15) a. The gift from the alumnus {matters / *matter}
b. The gift from the alumni {matters / *matter}
c. The gifts from the alumnus {*matters / matter}
d. The gifts from the alumni {*matters / matter}

(16) a. The alumnus's gift {matters / *matter}
b. The alumni's gift {matters / *matter}
c. The alumnus's gifts {*matters / matter}
d. The alumni's gifts {*matters / matter}

Both of these phenomena have (with good reason) been prominent test cases in work investigating connectionist systems' treatment of constituency-based generalizations. Studies using the question-forming rule as a probe into this issue include Frank and Mathis (2007), McCoy et al. (2020) and Warstadt and Bowman (2020), and those using subject-verb agreement include Linzen et al. (2016), Kuncoro et al. (2018) and Lakretz et al. (2021). And as illustrated in this section, the crucially constituency-sensitive rules underlying both of these phenomena bear on the distribution of English inflected verb forms (e.g. 'cleans' and 'cleaned') that Kallini et al. manipulate in order to create WordHop and NoHop. But English sentences with those inflected verb forms are a shared "starting point" for these two artificial languages, which differ only in whether the [S] and [P] markers occur in the hierarchically-defined position (iii) or at a count-based offset from *that* position. The constituency-based patterns in which verbal inflections participate — the dependency between the three positions illustrated in (11), and the hierarchical determination of the controller of agreement illustrated in (15)–(16) — are irrelevant for any *comparison* between WordHop and NoHop. WordHop contains just as much constituency-based question-formation, and just as much hierarchically-sensitive agreement, as NoHop does. A comparison between the two just amounts to a comparison between the count-based displacement in WordHop, and the absence of any analogous displacement in NoHop.

3 Towards a better comparison: counting vs. constituency

The problem with the comparison between WordHop and NoHop is that the count-based rule in WordHop is not the counterpart of any constituency-based rule in NoHop. There are two ways we might seek

[4]Relevant examples here are restricted by the fact that, in most varieties of English, subject-auxiliary inversion only occurs in matrix clauses. In some varieties spoken in Ireland, for example, the same operation applies in embedded clauses, yielding examples like 'I wonder will he clean it?' (McCloskey, 1992, 2006; Henry, 1995).

to rectify this. The first is to keep WordHop as our representative count-based language, and introduce a constituency-based rule to be the necessary counterpart: compare WordHop, where the [S] and [P] markers are placed at a count-based offset from position (iii), against a new synthetic language where these markers are placed at a constituency-based offset from position (iii). The second possibility is to keep NoHop as our representative constituency-based language, and *replace* one of the constituency-based rules governing the placement of the [S] and [P] markers with a count-based rule. Either route leads to some subtle issues that remain to be worked out, and my aim here is only to advance the discussion in a way that lays out the logic and clarifies what is needed, not to fully resolve the isssues that arise.

To introduce a constituency-based counterpart to WordHop's count-based rule, we need to identify a hierarchically-defined offset from position (iii) where markers would be placed in the new artificial language. Suppose we choose the right edge of the sister constituent of position (iii)'s parent V node; this will be the right edge of the direct object, in many cases. (There is no such constituent in the minimal diagrams in (11), but notice the relevant NP constituents in (6).) Then we would have a comparison between the count-based language illustrated in (17) (unchanged from Kallini et al.'s WordHop) and the constituency-based language illustrated in (18).

(17) Count-based (unchanged from WordHop)
- a. He clean his very messy bookshelf [S]
- b. He clean the bookshelf with glee [S]
- c. He clean it with a big [S] red broom
- d. He clean the bookshelf that is [S] messy

(18) Constituency-based
- a. He clean his very messy bookshelf [S]
- b. He clean the bookshelf [S] with glee
- c. He clean it [S] with a big red broom
- d. He clean the bookshelf that is messy [S]

One challenge here is that synthesizing examples like those in (18) from their English equivalents requires settling on an analysis of what counts as a sister of the relevant V node, which will be controversial in some cases; and in practical terms, even to the extent that analyses of individual cases are uncontroversial, they would likely to difficult to automate.[5]

A more subtle concern is whether the pattern in (18) is necessarily describable only in terms of a constituency-based offset from position (iii), or whether it has an alternative characterization in terms of a constituency-based offset from position (ii). If the position of the markers in (18) can be understood as part of a hierarchically-defined dependency with position (ii), then the pattern in (18) would be no better than NoHop: the comparison between (17) and (18) would again be a comparison between the composition of a constituency-based and a count-based offset from position (ii), and an only constituency-based offset from position (ii). The underlying question here is whether the composition of two constituency-based relations is always another valid constituency-based relation. The answer will depend on the details of one's theory of linguistically possible dependencies, which remains an active research topic. (It may bear repeating here that the exclusion of count-based dependencies is not one of the points of disagreement.)

Consider now the other route, where we pit a count-based rule against one of the existing constituency-based rules underlying NoHop. Let's suppose the relevant count-based rule placed the [S] and [P] markers at a four-word offset from position (ii), as a counterpart to the standard hierarchically-defined relationship between position (ii) and position (iii) illustrated in (11). This comparison is illustrated in (19) and (20).

(19) Count-based
- a. He clean his messy bookshelf [S]
- b. He always clean his messy [S] bookshelf
- c. He without doubt clean it [S]
- d. He clean it with a [S] broom

(20) Constituency-based (unchanged from NoHop)
- a. He clean [S] his messy bookshelf
- b. He always clean [S] his messy bookshelf
- c. He without doubt clean [S] it
- d. He clean [S] it with a broom

[5] Under any reasonable assumptions there will be many English examples where no such sister constituent exists, and these would need to be excluded from the corpora to keep the comparison balanced — just as Kallini et al. excluded sentences where an inflected verb was too close to the right edge for their WordHop rule to apply.

The necessary syntactic analysis for synthesizing examples like (19) only requires identifying position (ii) (the position of auxiliary verbs in standard declaratives), which is likely less controversial than the issues that arise for (18) surrounding sister constituents of the verb.

A questionable aspect of the comparison between (19) and (20) is that, in the constituency-based pattern, the word immediately preceding the marker is always of the same category (namely a verb), whereas in the count-based pattern the words preceding the marker are heterogeneous in syntactic category. (This is also a characteristic of the comparison between WordHop and NoHop.) This could be thought to make the constituency-based pattern more "predictable" or "simple" than the count-based one in a sense that we would like to control for. Notice that this consistency of an adjacent category is not a general property of constituency-based rules: in (18) we see the marker follow 'bookshelf', 'it' and 'messy', which belong to distinct syntactic categories. Rather it is a consequence of the fact that the rule that relates position (ii) and position (iii) in English ("affix hopping") is somewhat anomalous in ways that lead to divided opinions over whether it is best considered a morphological or syntactic rule (e.g. Halle and Marantz, 1993, pp.134–138; Embick and Noyer, 2001, pp.584–591).

As mentioned above, I make no attempt to resolve all these issues here; the main goal of presenting (17) vs. (18) and (19) vs. (20) is to lay out the logic of what would make an informative comparison between count-based and constituency-based rules, and in doing so clarify the earlier critiques of the comparison that Kallini et al. report.

4 Conclusion

In natural languages, words that are linked by some grammatical dependency do not always appear adjacent to each other. What linguists have taken to be striking is that the rules governing these non-adjacent configurations of co-dependent words are never describable in terms of (relative) numerical positions in the string; instead, the positions involved are characterized in constituency-based terms. This is hypothesized to be a consequence of an important difference in the status of count-based versus constituency-based rules in the human mind. Kallini et al. present their comparison between WordHop and NoHop as a test of whether GPT-2 shows an analogous asymmetry, but these two artificial languages do not differ in the appropriate way for this interpretation: the count-based rule in WordHop has no counterpart (constituency-based or otherwise) in NoHop, and so differences in learning success reflect the presence of this additional rule, not an asymmetry between two kinds of rules.

Of course, nothing I have said amounts to any claim about the underlying question of whether an LLM might exhibit a human-like asymmetry between count-based and constituency-based rules. The claim here is just that the experiments reported by Kallini et al. leave the issue untouched.

References

Akmajian, A., Demers, R. A., Farmer, A. K., and Harnish, R. M. (2001). *Linguistics: An Introduction to Language and Communication*. MIT Press, 5th edition.

Carnie, A. (2007). *Syntax: A Generative Introduction*. Blackwell, Malden, MA, second edition.

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

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Chomsky, N. (1971). *Problems of Knowledge and Freedom*. The New Press, New York.

Chomsky, N. (1975). *Reflections on Language*. Pantheon Books, New York.

Chomsky, N. (1980). On Cognitive Structures and Their Development: A Reply to Piaget. In Piattelli-Palmarini, M., editor, *Language Learning and Development: The Debate between Jean Piaget and Noam Chomsky*, pages 35–52. Harvard University Press, Cambridge, MA.

Chomsky, N. (1988). *Language and Problems of Knowledge*. MIT Press, Cambridge, MA.

Embick, D. and Noyer, R. (2001). Movement operations after syntax. *Linguistic Inquiry*, 32(4):555–595.

Frank, R. and Mathis, D. (2007). Transformational networks. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*.

Freidin, R. (1991). Linguistic theory and language acquisition: A note on structure-dependence. *Behavioral and Brain Sciences*, 14(4):618–619.

Freidin, R. (1992). *Foundations of Generative Syntax*. MIT Press, Cambridge, MA.

Fromkin, V., Curtiss, S., Hayes, B. P., Hyams, N., Keating, P. A., Koopman, H., Munro, P., Sportiche, D., Stabler, E. P., Steriade, D., Stowell, T., and Szabolsci, A. (2000). *Linguistics: An Introduction to Linguistic Theory*. Blackwell.

Halle, M. and Marantz, A. (1993). Distributed morphology and the pieces of inflection. In Hale, K. and Keyser, S. J., editors, *The View from Building 20*, pages 111–176. MIT Press, Cambridge, MA.

Henry, A. (1995). *Belfast English and Standard English: Dialect variation and parameter setting*. Oxford University Press, Oxford.

Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., and Potts, C. (2024). Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 14691–14714. `https://aclanthology.org/2024.acl-long.787`.

Kuncoro, A., Dyer, C., Hale, J., Yogatama, D., Clark, S., and 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*, pages 1426–1436.

Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., and Dehaene, S. (2021). Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699.

Lasnik, H., Depiante, M., and Stepanov, A. (2000). *Syntactic Structures Revisited*. MIT Press.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

McCloskey, J. (1992). Adjunction, selection, and embedded verb second. University of California, Santa Cruz.

McCloskey, J. (2006). Questions and Questioning in a Local English. In Zanuttini, R., Campos, H., Herburger, E., and Portner, P. H., editors, *Crosslinguistic Research in Syntax and Semantics*, pages 87–126. Georgetown University Press.

McCoy, R. T., Frank, R., and 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.

Radford, A. (1988). *Transformational Grammar: A First Course*. Cambridge University Press, Cambridge.

Warstadt, A. and Bowman, S. R. (2020). Can neural networks acquire a structural bias from raw linguistic data? In *Proceedings of the Annual Meeting of the Cognitive Science Society*.