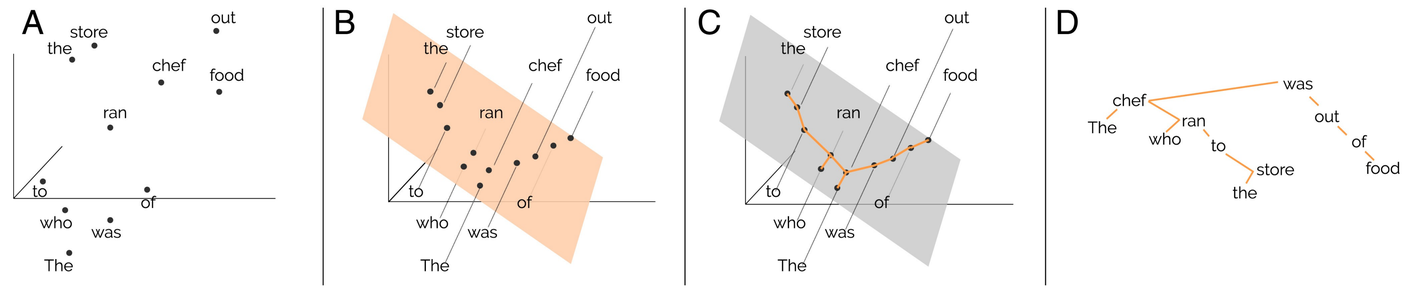
Evidence of Hierarchically-Complex Syntactic Structure Within BERT's Word Representations

Mary "Katie" Kennedy

1 Introduction

- Probing methodology seeks to investigate what linguistic phenomena a language model has encoded within its latent features (e.g., embeddings, attentions, etc.).
- Hewitt & Manning (2019) found LLMs have encoded enough syntax to **recover dependency trees** (Fig 1) wherein a head and dependent have a distance of ~1.
- But dependency trees are **shallow representations** that fail to capture deeper, more complex syntactic relations, like those that result from movement (Figs 3a–3b).

Fig 1. *Visualization of the probing method by Manning et al (2019).*



Each of the words of the sentence *The chef who ran to the store was out of food* is internally represented in context as a vector.

A structural probe finds a linear transform of that space under which squared L_2 distance between vectors best reconstructs tree path distance between words.

Once in this latent space, the structure of the tree is globally represented by the geometry of the vector space, meaning words that are close in the space are close in the tree.

In fact, the tree can be approximately recovered by taking a minimum spanning tree in the latent syntax space.

- **RQ: Are LLMs capable of capturing the hierarchical distances that are postulated in generative frameworks, like Minimalism?**

2 Methodology

We utilize Hewitt & Manning's (2019) dependency probe to train a linear transformation matrix to project a sentence's word representations into a subspace where a minimum spanning tree on the squared Euclidean distances recovers the dependency tree (Fig 1).

Our novel work deploys this dependency probe on structures whose ***dependency* parses are *identical***, but whose ***Minimalist* structures *differ***.

- *Why use a dependency probe when testing for generative syntactic structures?*
  - Constituency probes usually train on the English Penn Treebank, which is annotated with relatively atheoretical and "skeletal" syntactic structure
  - The edges between words varies more depending on the investigated structure

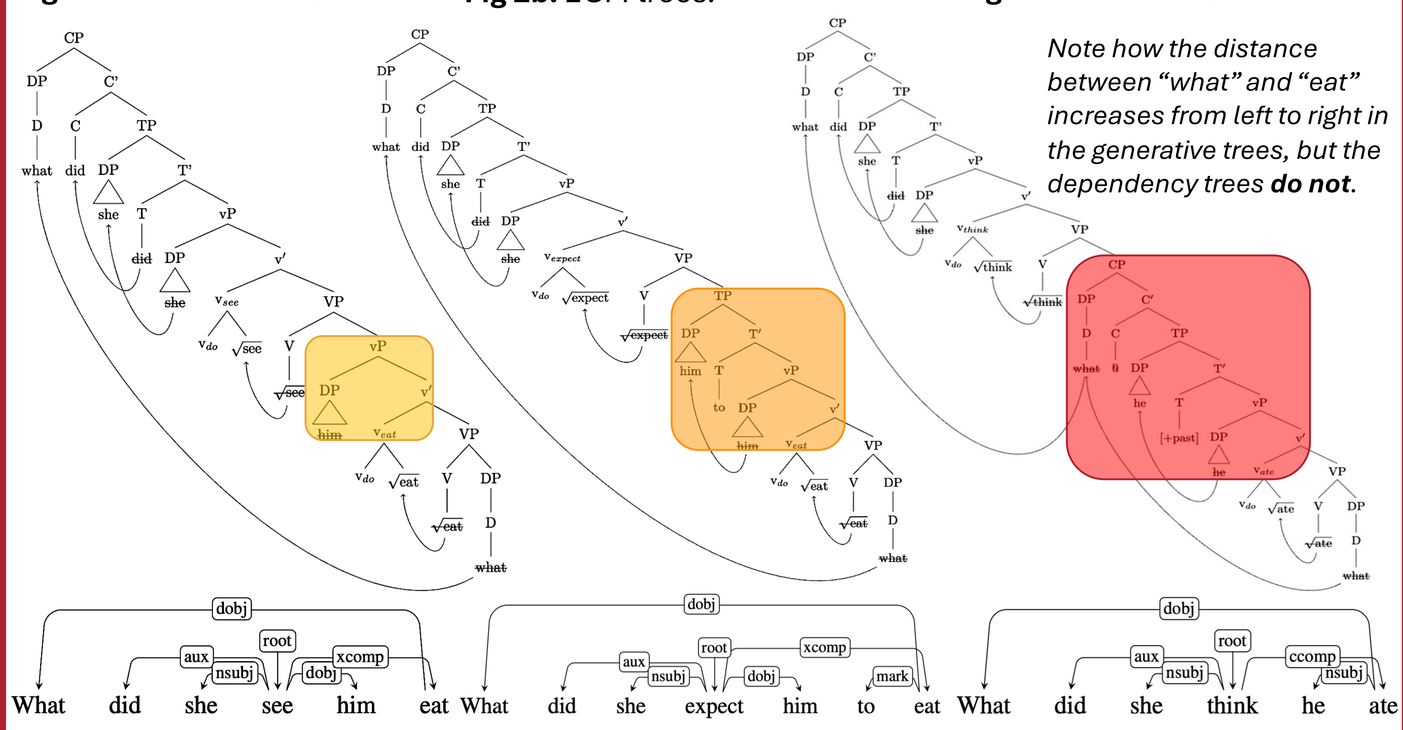
3 Design & Predictions

To probe whether contextualized vector embeddings encode Minimalist hierarchical distances, we use the filler-gap dependencies in *wh*-questions with embedded sentential complements that vary in size, such as:

- **Bare Infinitive** complements: *What did you see [$_{vP}$ him eat ~~what~~]*
- **ECM** complements: *What did you require [$_{TP}$ him to eat ~~what~~]*
- **Full CP** complements: *What did you think [$_{CP}$ he ate ~~what~~]*
- **Double ECMs**: *What did you expect [$_{TP}$ her to require [$_{TP}$ him to eat ~~what~~]]*
- **Double CPs**: *What did you believe [$_{CP}$ she suspected [$_{CP}$ he ate ~~what~~]]*

By varying the size of the embedded complement, the hierarchical distance between the extracted *wh*-word and its embedded verb **varies in a Minimalist** account, but ***not* in a dependency** account (see Figs 2a–b).

Fig 2a. *Bare Infin trees.* **Fig 2b.** *ECM trees.* **Fig 2c.** *Full CP trees.*



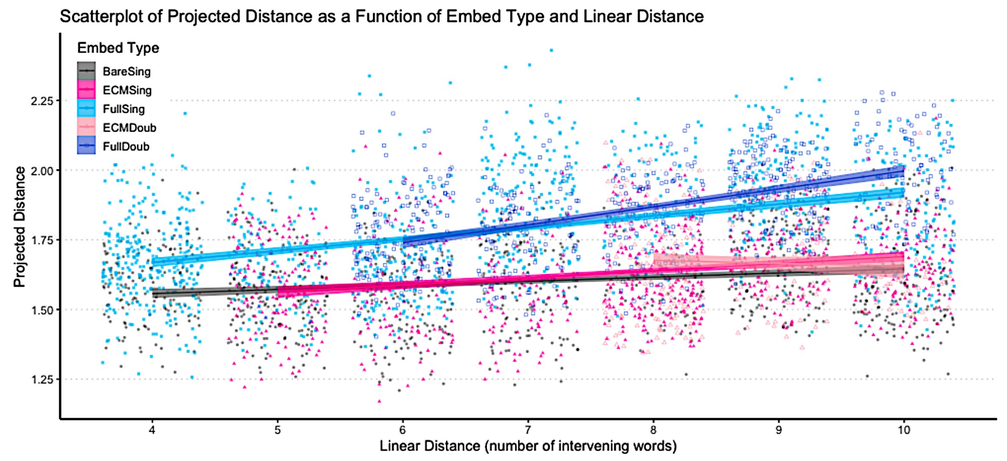
*Note how the distance between "what" and "eat" increases from left to right in the generative trees, but the dependency trees **do not**.*

If the embedding representations capture *only* head-dependency relationships and/or the probe is only sensitive to dependency grammar's shallow syntax, then **the probe's projected distance (~1) should *not* vary.**

However, if the representations capture Minimalist syntax, then **the projected distances should *increase* as the complement size increases.** Otherwise, it will remain unclear if LLMs either don't encode this hierarchical distinction or the dependency probe is simply not sensitive to these encoded differences.

4 Results & Discussion

Analysis. We conduct statistical analyses on the probe's projected distance ***iff*** the probe properly establishes a head-dependent relationship between the *wh*-word and the embedded verb (4,034 sentences of 18,2252 sentences).



Results. When linear distance is taken into account, the smallest BareSing complement is significantly **shorter** than any other complement size. Similarly, ECM is significantly **longer** than the *BareVP* and **shorter** than *FullCP* when analyzing only the singular clausal embeddings.

Discussion. The probe displays evidence of structural differences between the clause sizes, despite the dependency structures not varying. However, the rarity of double-clause embeddings in training decreases accuracy, which may obscure further findings on these more complex structures.

5 Conclusion

Findings.

1. Our finds suggest pretrained models like BERT have learned representations that to some degree approximate the hierarchical distinction between complement sizes.
2. A probe trained only to recover dependencies shows a sensitivity corresponding to a constituency-based analysis → Dependency Grammars may need to therefore consider accounting for complement sizes within their theoretical framework.

Limitations. Future exploration of this work would benefit from including extraction from multiple-clause embeddings within the training data to improve accuracy.

6 Selected References

1. John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. **2**. Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2019. Emergent linguistic structure in artificial neural networks trained by self-supervision. In *PNAS*.