Linguistic Minimal Pairs Elicit Linguistic Similarity in Large Language Models

Xinyu Zhou[1]* **Delong Chen**[2]* **Samuel Cahyawijaya**[2] **Xufeng Duan**[1] **Zhenguang G. Cai**[1]
[1]Department of Linguistics and Modern Languages, CUHK
[2]Department of Electronic and Computer Engineering, HKUST
xinyuzhou314@gmail.com, {delong.chen, scahyawijaya}@connect.ust.hk,
xufeng.duan@link.cuhk.edu.hk, zhenguangcai@cuhk.edu.hk

Abstract

We introduce a novel analysis that leverages linguistic minimal pairs to probe the internal linguistic representations of Large Language Models (LLMs). By measuring the similarity between LLM activation differences across minimal pairs, we quantify the *linguistic similarity* and gain insight into the linguistic knowledge captured by LLMs. Our large-scale experiments, spanning 100+ LLMs and 150k minimal pairs in three languages, reveal properties of linguistic similarity from four key aspects: consistency across LLMs, relation to theoretical categorizations, dependency to semantic context, and cross-lingual alignment of relevant phenomena. Our findings suggest that 1) linguistic similarity is significantly influenced by training data exposure, leading to higher cross-LLM agreement in higher-resource languages. 2) Linguistic similarity strongly aligns with fine-grained theoretical linguistic categories but weakly with broader ones. 3) Linguistic similarity shows a weak correlation with semantic similarity, showing its context-dependent nature. 4) LLMs exhibit limited cross-lingual alignment in their understanding of relevant linguistic phenomena. This work demonstrates the potential of minimal pairs as a window into the neural representations of language in LLMs, shedding light on the relationship between LLMs and linguistic theory.

1 Introduction

The categorization of linguistic phenomena[1] based on their relevance has been a long-standing endeavor, dating back to Aristotle (Aristotle, 350 BC). This has led to the widely accepted theoretical linguistic consensus of a hierarchical categorization of language structure encompassing syntax, semantics, morphology, etc., which provides a structured way to understand the intricate nature of language, and allows linguists to investigate the interrelationships and commonalities among these linguistic domains (Comorovski, 2013; Li, 2004).

Alongside the theoretical discussions of linguistic phenomena, a growing body of research on *quantitative measurement of similarities* based on statistical modeling on large-scale corpora has been observed in computational linguistics. Examples include lexical similarity (Holman et al., 2011), syntactic similarity (Boghrati et al., 2018; Schoot et al., 2016), semantic similarity (Pennington et al., 2014; Reimers and Gurevych, 2019), among others. These examples showcase the possibilities of understanding the nature of language through purely statistical methodologies. However, there has been limited research on quantitatively measuring the relationships between different linguistic phenomena. Given that language is a complex system composed of numerous interrelated linguistic phenomena, addressing this gap could lead to a more comprehensive understanding of language structure and its underlying mechanisms.

In this work, we aim to uncover and analyze the internal linguistic knowledge of Large Language Models (LLMs) when presented with a wide range of linguistic phenomena. LLMs are large-scale unsupervised language learners without any prior linguistic knowledge, and have demonstrated human-level language capability, as evidenced by their leading performance on language understanding benchmarks and impressive language generation fluency (Zhao et al., 2023; Bang et al., 2023). More specifically, we are interested in how LLMs represent different linguistic phenomena, and whether linguistically similar phenomena are represented similarly in LLMs.

*Joint first authors. Xinyu Zhou is now affiliated with Université Paris Cité and Sorbonne Université.

[1]Linguistic phenomena refer to observable patterns or features in language use. For example, subject-verb agreement is a linguistic phenomenon where verbs must agree with subjects in number and person. An example would be: *"The dog barks"* (correct) instead of *"*The dog bark"* (incorrect).

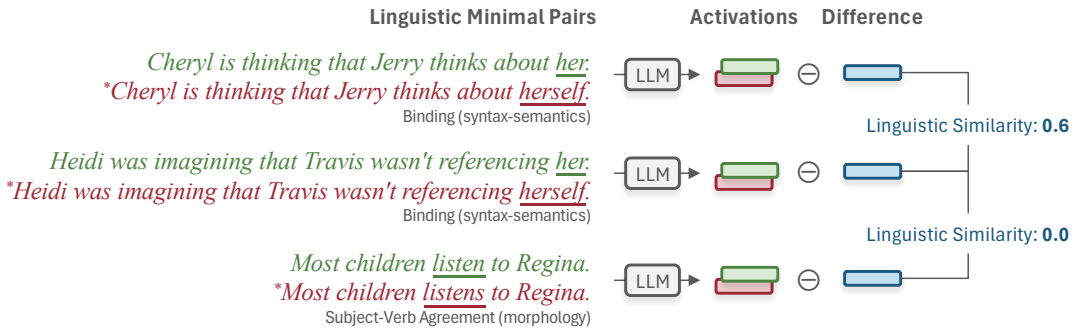

Figure 1: **The process of measuring linguistic similarity in an LLM**. We extract LLM activations for sentences in linguistic minimal pairs and compute their differences. Since the sentences differ solely in a specific linguistic phenomenon, the resulting difference only contains information about that phenomenon. We then measure the similarity between these activation differences, which we refer to as *linguistic similarity*.

To elicit such representations, we examine the activations in LLMs in response to linguistic minimal pairs (Warstadt et al., 2020). As shown in Fig. 1, these pairs consist of sentences that differ only in a word/phrase, with one being grammatical and the other ungrammatical. Since minimal pairs differ *only* in one particular linguistic phenomenon, information about other aspects (such as topic and semantic meaning) will be canceled out through subtraction. We interpret the remaining differences as the LLMs' internal representation of a specific linguistic phenomenon. By calculating the similarity between multiple such representations, we derive a measure of *linguistic similarity* between linguistic minimal pairs.

We then conduct an extensive analysis of linguistic similarity in LLMs. Our experiment encompasses 100+ LLMs of varying scales and pretraining corpora, utilizing 150,000 linguistic minimal pairs across 3 different languages. We report our observations correspond to the following key questions:

1) How consistent is linguistic similarity across different LLMs? LLMs have the highest alignment of linguistic similarity in English, which is the most widely used language for LLM pertaining, while the alignments are comparatively weaker in Chinese and Russian. We further visualized the relationships among these LLMs with UMAP (McInnes et al., 2018). On Chinese samples, we observed a distinct clustering pattern: bilingual and multilingual LLMs formed one cluster, while English-only models formed another. The above results suggest that the language distribution in the training data influences the linguistic similarity in LLMs.

2) Does linguistic similarity align with theoretical linguistic categorizations? We compared linguistic similarity across three levels of theoretical linguistic categorizations. Our analysis revealed that fine-grained classifications exhibit significantly higher intra-class similarities compared to inter-class similarities. However, this disparity diminishes considerably at higher categorization levels. Meanwhile, we can also observe some highly correlated phenomena pairs that are not classified to the same theoretical categorization.

3) To what extent does linguistic similarity correlate with semantic similarity? We showed a weak correlation between semantic similarity and linguistic similarity, despite many existing samples with low linguistic similarity and high semantic similarity, and conversely, high in linguistic and low in semantic. The weak correlation indicates that linguistic similarity in LLMs has a *context-dependent* nature.

4) Whether relevant phenomena in different languages enjoy higher linguistic similarities? We compare the linguistic similarity of the shared three linguistic phenomena in English and Chinese. Our UMAP visualization revealed that while English phenomena are clustered within a shared region, they are "attracted" by their relevant phenomena in Chinese.

We hope this paper sparks new exploration into LLMs' internal linguistic representations, uncovering deeper insights into their inner workings and potentially informing linguistic theory. To facilitate future research, the activation differences of the 100+ LLMs, pre-computed sample-level linguistic similarities, and all the codes are made publicly available at `https://github.com/ChenDelong1999/Linguistic-Similarity`.

References

Aristotle. 350 BC. Categories.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718.

Reihane Boghrati, Joe Hoover, Kate M Johnson, Justin Garten, and Morteza Dehghani. 2018. Conversation level syntax similarity metric. *Behavior research methods*, 50:1055–1073.

Ileana Comorovski. 2013. *Interrogative phrases and the syntax-semantics interface*, volume 59. Springer Science & Business Media.

Eric W Holman, Cecil H Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, et al. 2011. Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, 52(6):841–875.

Yafei Li. 2004. *Xo: A theory of the morphology-syntax interface*. MIT Press.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Lotte Schoot, Evelien Heyselaar, Peter Hagoort, and Katrien Segaert. 2016. Does syntactic alignment effectively influence how speakers are perceived by their conversation partner? *PloS one*, 11(4):e0153521.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.