Emergence of Phonemic, Syntactic, and Semantic Representations in Artificial Neural Networks

Anonymous Authors[1]

Abstract

During language acquisition, children successively learn to categorize phonemes, identify words, and combine them with syntax to form new meaning. While the development of this behavior is well characterized, we still lack a unifying computational framework to explain its underlying neural representations. Here, we investigate whether and when phonemic, lexical and syntactic representations emerge in the activations of artificial neural networks during their training. Our results show that both speech- and text-based models follow a sequence of learning stages: during training, their neural activations successively build subspaces, where the geometry of the neural activations represents phonemic, lexical, and syntactic structure. While this developmental trajectory qualitatively relates to children's, it is quantitatively different: These algorithms indeed require 2 to 4 orders of magnitude more data for these neural representations to emerge. Together, these results show conditions under which major stages of language acquisition spontaneously emerge, and hence delineate a promising path to understand the computations underpinning language acquisition.

1. Introduction

To acquire language, humans follow a remarkably stereotyped developmental trajectory: infants first learn to distinguish phonological categories (Kohl, 1993; Eilers et al., 1979; Jusczyk et al., 1993; Kuhl, 2004), and only later become able to understand and produce complex semantic and syntactic structures (Gleitman, 1990; Pinker, 1996; Christophe et al., 2008; Gomez & Gerken, 1999; Friedmann & Rusou, 2015; Boudreault & Mayberry, 2006). Linguistic theory has long posited the existence of distinct structures for these various domains of language – for example, nested-trees for syntax, hierarchical graphs for lexical semantics, and the vowel trapezoid for phonemes. While the acquisition of such structures is increasingly well characterized (Friedmann et al., 2021; Dupoux, 2018), the underlying neural and computational mechanisms of language acquisition remain poorly understood (Kuhl, 2010; Frank, 2011; Martin et al., 2022; Kachergis et al., 2022; Dubois et al., 2016; Evanson et al., 2025).

Correlates of these structures were found to organize large language models (LLMs) activations, and consequently, LLMs appeared as a fruitful path to propose a mechanistic theory of language acquisition (Frank, 2023; Vong et al., 2024; Dupoux, 2018). Specifically, the activation patterns of self-supervised models represent simple semantic relationships (Landauer & Dumais, 1997; Jawahar et al., 2019; Pasad et al., 2021; 2024; Park et al., 2025), syntactic structure (Hewitt & Manning, 2019; Tenney et al., 2019; Evanson et al., 2023) and phonological categories (Baevski et al., 2020a; De Heer Kloots & Zuidema, 2024).

However, it is unknown whether modern LLMs learn to instantiate different linguistic structures using a common principle. We suggest that this gap of knowledge is due (1) to the opacity of neural networks whose representations instantiate simultaneously all linguistic structures, and (2) to the lack of longitudinal measurement of these representations during training. To fill this gap, we study the geometry of these structures in neural space and the order in which they emerge in the models during training, much like in human language acquisition. We generalize (Hewitt & Manning, 2019)'s structural probe to test whether and when the representations of phonemes, lexical semantics and syntactic trees emerge during the training of speech and text models. More precisely, we test the hypothesis that these linguistic structures are neurally encoded using a geometric code. We evaluate whether there exists a subspace of activations in which distances between model activations linearly correspond to the phonemic, lexico-semantic, or syntactic distances predicted by linguistics.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

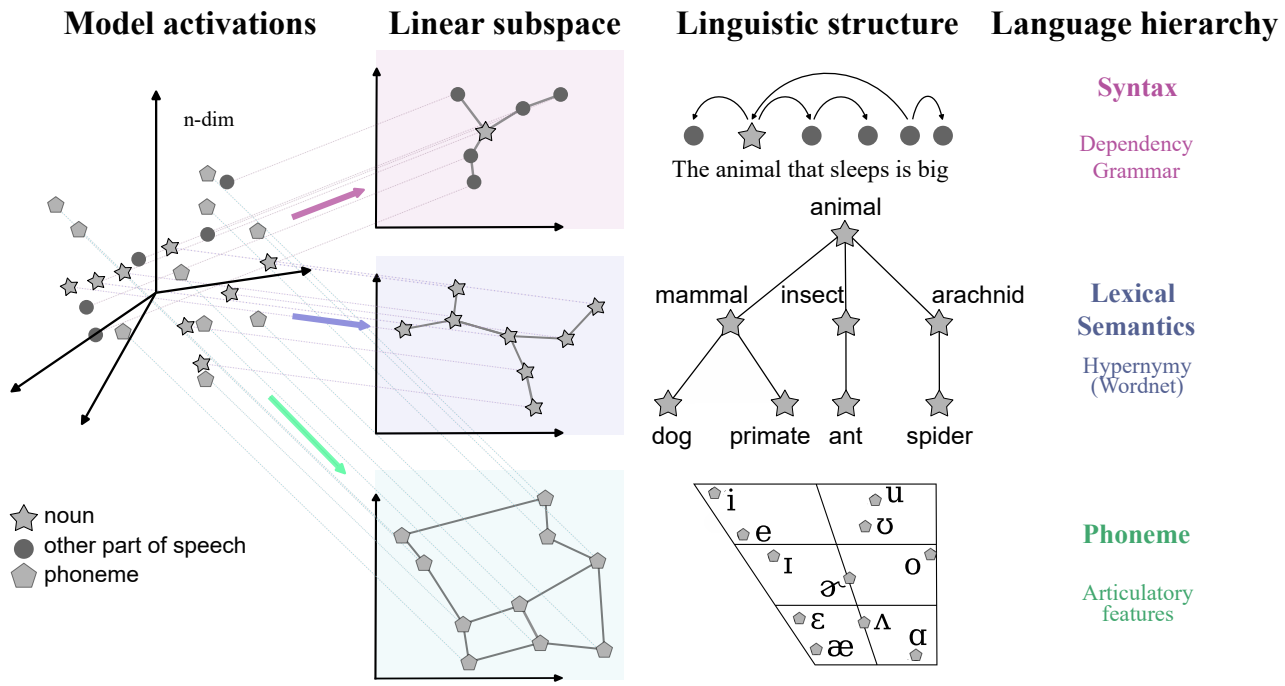

Figure 1. **Linear probing recovers phonemic, lexical semantic, and syntactic structure from the activation spaces of neural networks trained via Self-Supervised Learning.** The networks' neural activations live on a high-dimensional space in which different linear subspaces coexist. Different linear probes recover linear subspaces which represent phonemic, lexical semantic and syntactic structures, as long postulated by linguistic theories.

2. Methods

2.1. A Shared Linear Probe for extracting Linguistic Structures from Model Activations.

Problem. Linguistics predicts that language inputs (e.g. text, speech) should be represented with different levels of representations (e.g. phoneme, words, syntax). Each level of representation can be formalized as a metric system, which indicates e.g. whether different phonemes are more or less similar to one-another or whether different words in a sentence are more or less syntactically related. Formally, this prediction implies that language representations reduce down to a distance matrix. Here, we aim to evaluate whether language models learn to build subspaces in which activation pattern distances correlate with these distance matrices. For this, we extend the use of the *Structural Probe* (Hewitt & Manning, 2019), originally restricted to syntax, to any language representation, and fit this linear probe to identify subspaces that represent phonemic, lexical semantic, and syntactic features. Following this established interpretability paradigm, we restrict our focus to probing of representations can be directly used by downstream neurons through a simple linear readout. In contrast, non-linear probing could recover the target distances from arbitrary representations, thus hindering interpretability.

Objective. Consider a set of activation vectors of dimension k in response to n elements (e.g. word, token, phonemes): $h \in \mathbb{R}^{n,k}$. Given a pairs of elements (i, j) in the set of possible pairs S, and a target distance $d_{i,j} \in \mathbb{R}^+$, the *Structural Probe* aims to find a linear transformation $B \in \mathbb{R}^{k,p}$ to extract the subspace where the distance between the elements' activations predict their target distance (Hewitt & Manning, 2019).

$$\hat{B} = argmin_{B \in \mathbb{R}^{k,p}} \sum_{(i,j) \in S} |||(h_i - h_j)B||_2^2 - d_{i,k}|$$

Following Hewitt & Manning (2019), we use the squared Euclidean $|| \cdot - \cdot ||_2^2$ distance in the projected space. We optimize 2D ($B \in \mathbb{R}^{k,2}$) probe for visualization and 200D ($B \in \mathbb{R}^{k,200}$) probe for evaluation.

Evaluation. For consistency with previous works (Hewitt & Manning, 2019; Diego Simon et al., 2024; Limisiewicz & Mareček, 2021), we evaluate a probe on each level of representation (Phoneme, Lexical and Syntax) using a Spearman correlation between true distances (proposed by linguistics) and those measured in the probe space.

2.2. Probing Datasets

We construct a paired text-speech benchmark made of different text datasets and their artificially synthesized speech

counterparts (Table A in appendix).

Syntactic representations. The Universal Dependencies English Web Tree Bank (UD–EWT) corpus (Silveira et al., 2014) contains 16,622 sentences extracted from different web media. For each sentence, the syntactic tree is manually annotated by linguists according to the Universal Dependencies formalism (de Marneffe et al., 2021; Nivre et al., 2016), a type of Dependency Grammar (Tesnière, 1959). The UD–EWT corpus has been widely adopted to train both syntactic probes (Müller-Eberstein et al., 2022; Diego-Simón et al., 2024; Limisiewicz & Mareček, 2021) and syntactic parsers (Grünewald et al., 2021; Hershcovich et al., 2018; Yuan et al., 2019). For consistency, we keep the original train, validation and test splits provided in UD–EWT. The golden distance between two words of a sentence was defined as the number of edges between each word in the dependency tree of this sentence.

Lexical representations. Lexical semantics is assessed as a hypernymy relationship (is a generic term of) between units of senses (synsets) associated to nouns, in the WordNet semantic graph. For example, the synset "animal.n.01" is a hypernym to "cat.n.01". To obtain a maximal coverage of WordNet nouns, we construct a dedicated word-level dataset, named *WordNet–Nouns*, that contains most noun synsets in WordNet, while avoiding major polysemic issues. This procedure is detailed in the appendix. We focused on the noun subgraph of WordNet because including several parts of speech (POS) categories can lead to a large semantic score just because of an ability to discriminate between POS, and so, mislead to the conclusion that probes have learnt semantic structure. This resulted in a total of 35,352 sets of words. For the speech modality, we construct *WordNet–Nouns–TTS* by synthesizing every word with `seamless-m4t-v2-large` (Barrault et al., 2025) and aligning the resulting waveforms with the Montreal Forced Aligner (McAuliffe et al., 2017). To define a semantic distance, we turned to path length along the WordNet gaph. The gold distance between synsets A and B counts the number of WordNet edges (hypernym/hyponym relations) linking one synset to another.

Phonemic representations. To create a speech counterpart to UD–EWT, we synthesize all the sentences in the UD–EWT corpus using the Text-To-Speech model `seamless-m4t-v2-large` (Barrault et al., 2025). We then perform forced alignment with the Montreal Forced Aligner (McAuliffe et al., 2017) to obtain word and phone-level time-stamps for every synthesized utterance. Phoneme symbols were converted from ARPABET to IPA using the panphon library (Mortensen et al., 2016). Consequently, the resulting speech corpus preserves the original UD–EWT syntactic annotations while enriching them with phoneme identity labels. To define the distances among the different English vowel phonemes, we used dissimilarities of articulatory feature vectors from (Mortensen et al., 2016). The dissimilarity value between 2 phonemes is an integer ranging from 0 (most similar) to 12 (most different), which represents the number of non-shared articulatory features.

2.3. Models

Text models. We use large language models from the LLama2 and Pythia suite (Biderman et al., 2023). Pythia models were auto-regressively trained on English text (Gao et al., 2020), differing only in their parameter size. Within this suite, we use models ranging in size from 70M to 1.4B parameters, for which the intermediate training checkpoints are publicly available [1]. The two Llama 2 models, of size 7B and 13B, allowed us to score models with stronger linguistic scores (Waldis et al., 2024), defining an upper-bound for what the smaller Pythia models could achieve.

Speech models. We use audio Wav2Vec 2.0 models (Baevski et al., 2020b) previously trained on English speech, French speech, environmental sounds, and music (Orhan et al., 2025; Parcollet et al., 2024). Wav2Vec 2.0 models are pretrained in a self-supervised fashion to reconstruct masked segments of their own latent speech representations, an approach that encourages context-aware acoustic features. However, it remains unclear whether this self-supervised objective also fosters higher-level syntactic and semantic capabilities.

3. Results

3.1. Structural probes recover the geometry of phonemic articulation.

Phonemes can be discriminated from the activations of self-supervised speech models (Baevski et al., 2020b; Ji et al., 2022; Seyssel et al., 2022; De Heer Kloots & Zuidema, 2024). It remains unclear if their code for phonemes is structured according to articulatory features. To measure this, structural probes are trained to identify a phoneme subspace that instantiates articulatory features.

Visualization and scores across models' layers We first visualized the representations of a Wav2Vec 2.0 model pretrained on English speech through a less powerful 2-dimensional probe, optimized on the UD-EWT-TTS dataset. Remarkably, the projection of the test set (Fig. 2A) represents vowels similarly to the classic vowel trapezium (Fig. 2B): vowel phonemes with a similar articulation tend to be closer in the phonemic subspace. To evaluate how well the model instantiates the phoneme structure, we then train

[1] https://huggingface.co/models?other=pythia

probes with a higher dimensionality. We find that phonemes are best encoded in layers with a relative depth of 0.5 and 0.8 (Fig. 2C), achieving a Spearman correlation of 0.75. This phonemic score outperforms by a large margin the untrained, random, baseline (phonemic score of 0.2). This demonstrates that instantiating the phoneme structure is not performed trivially with random acoustic processing.

Effect of model size. Is this a pure effect of model training, or is it also dependent on the model size? We find that the phonemic score scales with model size: larger models achieve higher scores than smaller ones (Fig. 2D). Among English-trained models, Wav2Vec 2.0 large (315M parameters) outperforms base (94M), which in turn outperforms tiny (30M), each models implemented with 24, 12, and 3 Transformer layers, respectively. This highlights that, despite its apparent simplicity, extracting the phoneme structure is easier with more contextual processing layers.

Controls: acoustic pretraining is not enough To verify that the instantiation of phoneme structure was due to speech processing instead of generic acoustic processing, we compared models trained on different data. Phonemic scores depend strongly on the training data (Fig. 2D): models trained on English speech perform best, whereas those trained on French, music, or environmental sounds perform poorly. Note also, that the music and environmental sounds models are not completely naive to speech, with an estimated 8% of speech inputs in the environmental sound dataset (Orhan et al., 2025). Only the French large and base models outperform the `tiny` English model, which is expected given the model's limited size. These controls prove that the instantiation of phoneme structure results from the emergence of speech-specific processes.

Emergence during pretraining How is this representation of phonemes learned? We probe the English Wav2Vec 2.0 94M model across its training checkpoints. For simplicity, we focus on the layer with the best phonemic score, i.e., layer 9. Fig. 2E shows a gradual rise in this score with the cumulative spoken-word count, reaching emergence after roughly 10^9 words (including repetitions), which corresponds to about 90,000 pretraining steps. This is considerably more than the 10^7 words that children experience in their first year on average (Gilkerson et al., 2017). Taken together, these results indicate that self-supervised speech models spontaneously learn to instantiate a phoneme subspace whose geometry mirrors distances of articulatory features.

3.2. Lexical Semantics

Language model representations have been known to encode semantic features (Korchinski et al., 2025). Additionally, models' behavior and the organization of their final layer have been shown to be consistent with an instantiation of lexical semantic relationship (Cohen et al., 2023), including the WordNet graph (Park et al., 2024). To measure the internal layers' representation, we train structural probes to predict the WordNet noun graph from text and speech models.

Visualization and scores across models' layers. We first optimized a 2-D probe of text models' semantic representations in order to visualize their representations. We display the training and test elements together while evaluating only across test elements. To our surprise, a 2-D probe already embeds well-constructed semantic graphs, especially when we focus on small subsets of the larger WordNet tree. As an example, we plot in Fig. 3A, a 2-D embedding of the tree below the mammal synset from the activations of a Pythia-1.4B text model. Other examples for several "supersense" semantic categories are available in Figs. E and F in appendix, as well as an example of 2-D embedding of the whole WordNet tree, see Fig. B in appendix. In two dimensions, this embedding does not capture the exact tree, but rather positions well the largest category of this semantic tree, as illustrated by plotting these categories' centroid in Fig. 3A-B (diamond markers). Smaller categories and the exact edges of the tree require a larger embedding space. We consequently train and evaluate a 200-D probe of the whole WordNet graph, in text and audio models. For this 200-D probe, the semantic scores across model layers follow an inverted U-shape in all models: Wav2vec 2.0 (160M, max 0.1), Pythia (1.4B, max 0.29), and Llama2 (13B, max 0.41), Fig. 3C. This is remarkable, because transformer's tokens are vectorised and could therefore contain some semantic information. Here, we show that such semantic information is not sufficient to instantiate the lexical semantic structure of nouns. Instead, several computational steps are performed to merge and position these tokens' embedding together into a semantic subspace.

Effect of model size. This questions the computational power necessary for these semantic spaces to emerge. We measured the maximal semantic score across layers as a function of parameters in a set of text and audio models, Fig. 3D. Remarkably, semantic scores increase linearly with text model sizes in the Pythia model suite and finally plateau at a large (7B) size for the Llama2 models. Similarly, the semantic scores of audio models increase with model size, although through a less steep slope. At equivalent size, text models have slightly larger semantic scores than audio models. This is in stark contrast with their very distinct training regime, with audio models seeing a completely distinct amount of data. This hints at a possible different training dynamic between audio and text models.

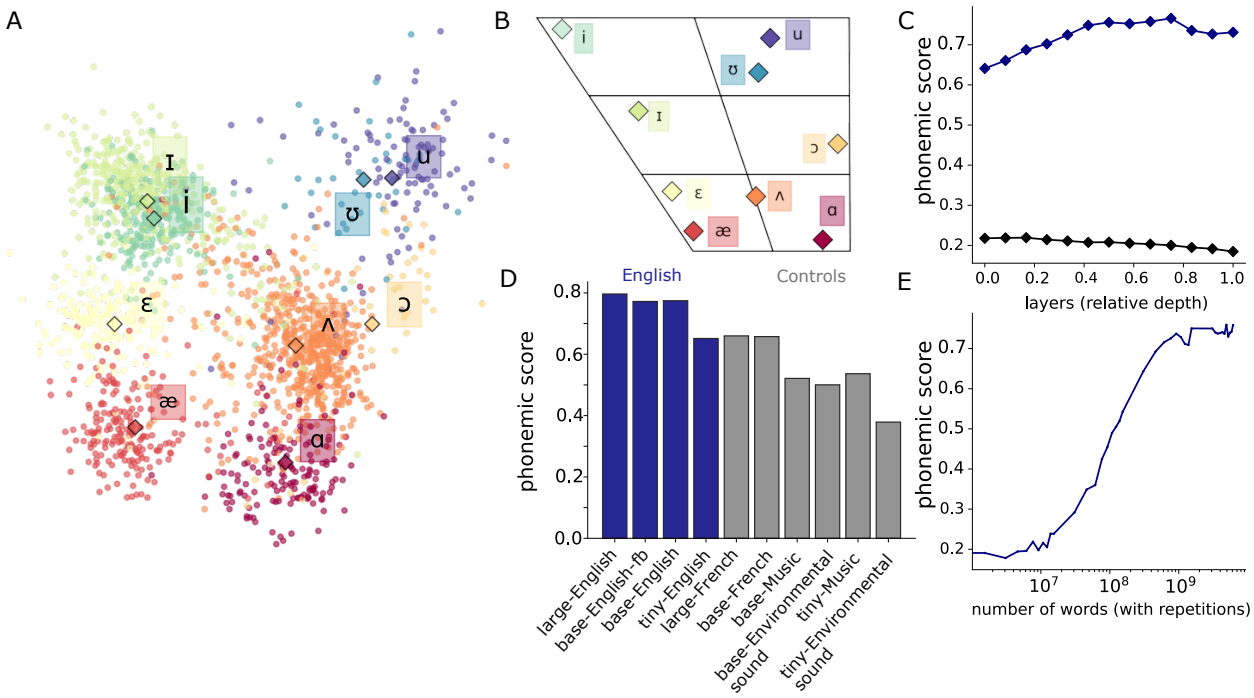

Figure 2. **Phoneme: structural probes recover the geometry of phonemic articulation. A.** Phoneme representations from a 2D structural probe trained to match pairwise articulatory distances among vowels. Responses to the same phoneme are colored identically. **B.** Canonical vowel trapezium indicating tongue height (vertical) and backness (horizontal). **C.** Layer-wise phonemic score for Wav2Vec2: probes trained on the final pretraining checkpoint peak near 80% relative depth and consistently outperform probes trained on the initial (untrained) checkpoint. **D.** Effect of model and data: English pretraining and larger model capacity both yield higher phonemic scores. **E.** Phonemic structure emerges gradually with pretraining, with scores increasing with the number of pretraining steps; the data required for Wav2Vec 2.0 to acquire robust phonemic structure far exceeds typical estimates of children's linguistic input.

Emergence during pretraining. We measured the evolution of the semantic score as a function of pretraining of a base Wav2vec 2.0 model (94M) trained on librispeech data, and a 1B Pythia model, Fig. 3E. Audio models' semantic scores initially increased faster than text models' scores, but then saturated. This saturation of the audio models questions whether they discovered semantic structure or if these scores could be explained by the acoustic properties of the words.

Controls: semantic mostly arise from acoustic cues in audio models Acoustic cues could be sufficient to encode a small part of the semantic subgraph. For example, "dog" and "hunting dog" are expected to be close by in the graph, but this could be predicted purely from acoustic features; here, the fact that "dog" is present in both sets of words. To control for these acoustic features, we evaluated tiny (30M) and base (94M) Wav2vec 2.0 models trained on music, environmental sounds, and a large (315M) model trained on French speech. Each of these audio models reached significant semantic scores, and these scores were in the same range as the semantic scores of models exposed to English speech, but remained much lower than the scores of text models (Table D in appendix). With this structural probe, it consequently seems that most of the audio models' scores are due to acoustic correlates. Despite their low semantic score, the English-exposed models better clustered 89% of semantic categories of the WordNet tree than control models, with a 0.1 F1-score gain on average (Fig C in appendix, paired t-test 3.8×10^{-36}). Together, these results demonstrate a clear and modest but significant instantiation of lexical semantic structure by text and audio models, respectively.

3.3. Syntax

Language models are known to code syntactic trees through distances (Hewitt & Manning, 2019), while syntactic trees per se have yet to be found in speech models, since previous approach focused on decoding statistical quantities derived from the trees, like node depth (Pasad et al., 2024). Here we train structural probes to measure if speech models also use a distance code to instantiate syntactic trees.

Visualization and scores across models' layers. We first visualized the 2-D probing of a Wav2Vec 2.0 base (94M) model activations in response to 6 sentences with increasing hierarchical depth (Fig. 4A and B). As expected, the true

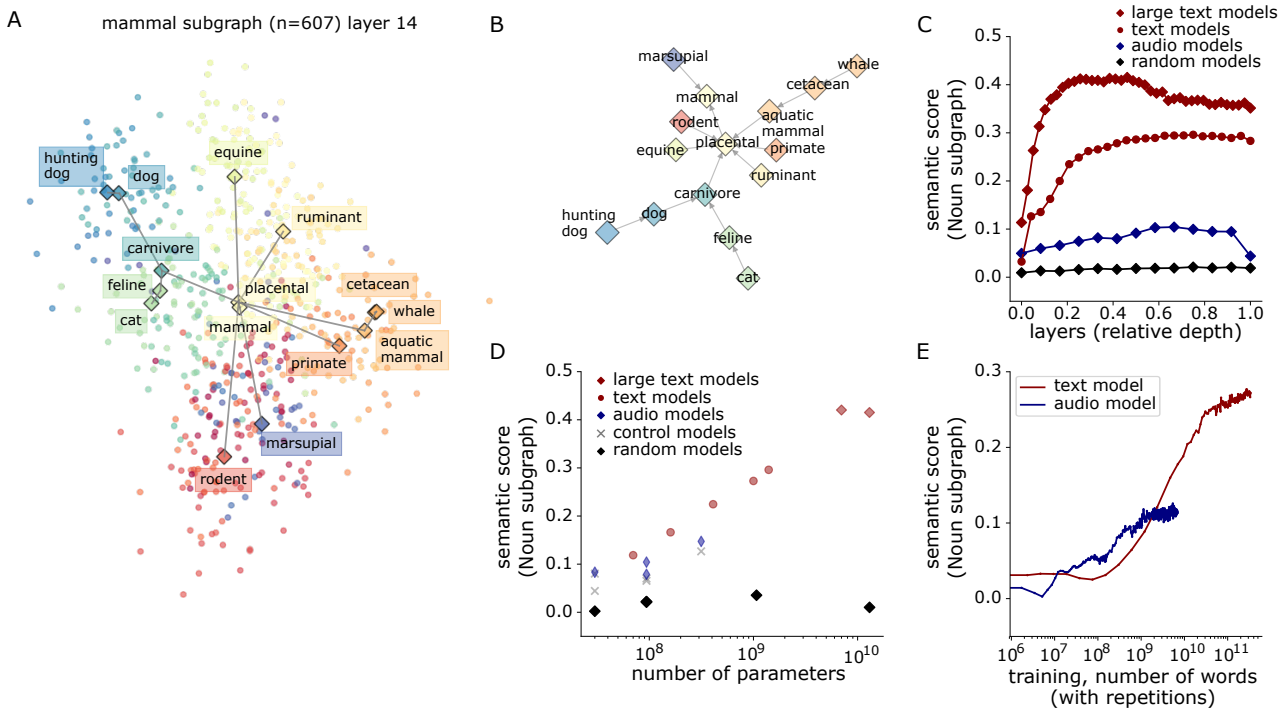

Figure 3. **Structural probes recover the geometry of lexical semantic graphs** A: 2d projection of text model (pythia-1.4B) response to all words in the mammal subgraph. B: Example WordNet subgraph, displaying synsets of the mammal subgraph with more than 50 hyponyms. c: Semantic score (for the graph composed of all nouns) for a large text (Llama-13B), text (pythia-1.4B), audio (Wav2vec 2.0-94M base), and random (Wav2Vec 2.0-94M base) models. D: Semantic scores for all models as a function of the model size. E: Semantic score as a function of the quantity of pertaining for the text (pythia-1.4B) and audio (Wav2vec 2.0-94M base) models.

(gold) and predicted (black) syntactic trees closely matched for simple sentences, but became increasingly more distant for harder, complex sentences. We then evaluated syntactic scores with probing in 200 dimensions. We observed strong syntactic scores for all models, including Wav2vec 2.0 models (94M parameters, max: 0.76), Pythia model (1.4B, max: 0.81), Llama2 model (13B, max: 0.82) (Fig. 4C).

Effect of model size For text and audio models, syntactic performances first increase then plateau as a function of the number of parameters, Fig. 4D. Remarkably, unlike semantic scores, text models' performances quickly plateau with model size. This indicates that syntactic structure computations do not require very large text models, but also that large text models do not develop a perfect syntactic parsing algorithm.

Emergence during pretraining We plot in Fig. 4E the evolution of the syntactic score as a function of pretraining. Unlike text models, audio models' scores did not seem to saturate with the quantity of pretraining. For models of equivalent size, the emergence of syntax was faster in audio models, for which the dataset also used a smaller quantity of different sentences. This suggests that audio data provides syntactic cues that the model can discover and use to build syntactic subspaces.

Controls: models instantiating syntax encode more than a linear tree Similarly to semantics, could syntax processing in audio models be explained by acoustic features? If indeed control audio models reached relatively high syntactic scores, we found that this came from the oversimplicity of the metric. Indeed, control audio models partially predicted the syntactic trees, purely from the fact that they could predict a linear tree (where words are linked to their neighbors). To demonstrate this, we computed the Spearman correlation with a linear tree and compared it to the correlation with the gold dependency tree. This correlation were similar in the case of control models but not for English-exposed models. Critically, this was not the case for English-exposed models whose projected representations correlated more with the syntactic than the linear structure. This demonstrates that true syntactic representations in response to English sentences are present in English-exposed models but not in models trained on French, music, or environmental sounds.

3.4. Order of acquisition

We next compare the learning dynamics of these phonemic, semantic, and syntactic codes in the audio model. To reveal

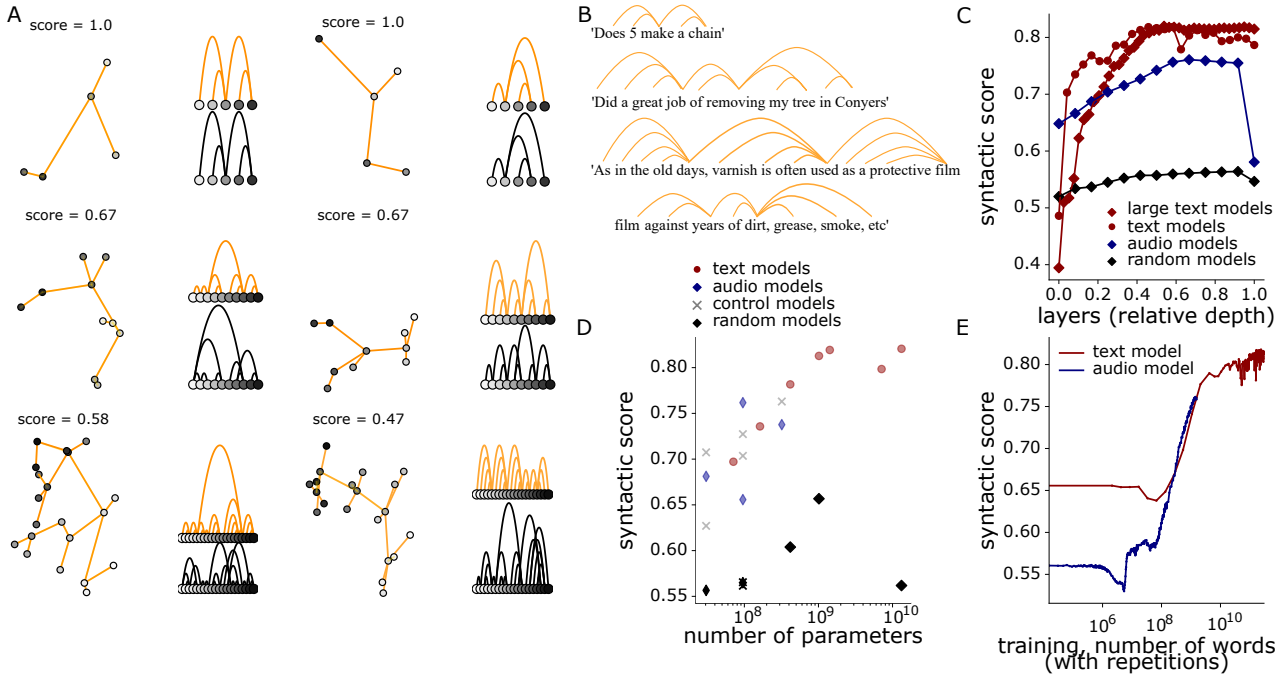

Figure 4. **Structural probes recover the geometry of syntactic trees** A. In each left subpanel, we plot an example of a 2d probe projection of the model activations, with trees reconstructed through the minimum spanning tree algorithm. On the right subpanel, we plot the gold (yellow) and predicted (black). B. Example of sentences and their syntactic trees. C. Syntactic scores of a large text (Llama-13B), a text (pythia-1B), speech (Wav2vec 2.0-94M base) and random (Wav2vec 2.0-94M base) models for each transformer layer. D. Highest syntactic score across layer for each model. E. Emergence of the syntactic abilities during pretraining, which includes several repetitions of the same dataset for the audio model.

these emergences, we repeated phonemic, syntactic, and semantic probing across checkpoints logarithmically spanning the pretraining of the models. We plot in Fig. GA-C in supplementary the probing of 3 checkpoints that highlight the successive emergence of phonemic and lexical-syntactic structures in the audio models. We plot in Fig. 5A the emergence of these scores and fit a parametric curve for each. This allows us to then plot in Fig. 5B a relative score for each curve and compare their dynamic. Remarkably, we observe that phonemic emergence is followed by a partial emergence of lexical semantic and finally an emergence of syntactic abilities in the audio models. These emergences are each roughly separated by an order of magnitude of input data, indicating that each stage solidifies before the previous one. Yet, as a semantic code only partially emerges in the audio model, it remains to be seen if these models truly predict that lexical semantic structure appears before syntactic structure.

4. Discussion

Goal and approach To evaluate whether modern language algorithms can help understand the computational underpinning of language acquisition, we generalized Hewitt & Manning (2019)'s Structural Probe and assessed the emergence of phonemic, lexical, and syntactic representations during the training audio and text models.

Similarities to Human Language Acquisition. In spite of their remarkably distinct architecture and learning principles, audio models appear, like children, to spontaneously build linguistic features, in a developmental trajectory that, at least at a coarse view, follows a phonemic and lexico-syntactic sequence. This emergence effectively captures some of the developmental trajectory of children, who – during their first year of life – demonstrate a progressive categorization of phonemes, and subsequently develop an ability to understand and produce syntactic structures from an increasingly large and precise vocabulary (Gomez & Gerken, 1999; Kuhl, 2004; Dupoux, 2018; Friedmann & Rusou, 2015; Boudreault & Mayberry, 2006; Sainburg et al., 2022).

A quantitative gap. Consistent with previous work (Dupoux, 2018; Lavechin et al., 2022; Frank, 2011; Hu et al., 2024), data efficiency remains a major gap: compared to human children, the presently-studied speech and text models require 2-4 orders of magnitude more input data for these representations to emerge (Fig. 5 F) (Gilkerson et al., 2017). This discrepancy highlights that modern lan-

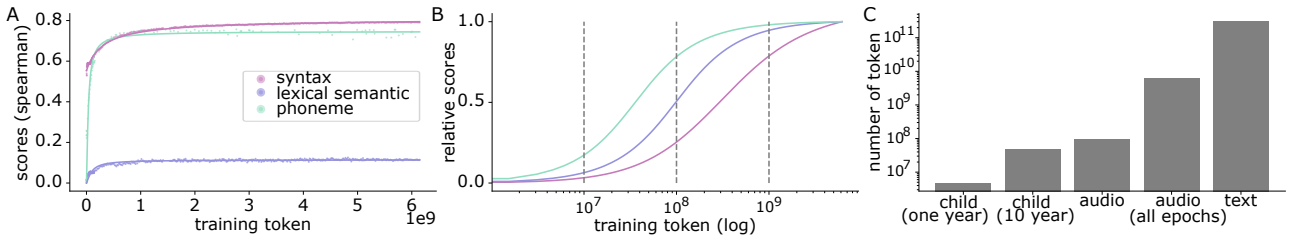

Figure 5. **Linguistic structures are acquired in sequential order** A: Left: Emergence of semantic, syntactic, and phonemic score for a Wav2vec 2.0-94M base model (dots), along with a parametric fit (line) B: relative scores of the parametric fit (min to 0 and max to 1), demonstrating the successive emergence of phonemic and lexical-syntactic structures. Checkpoints visible in Fig.G in the appendix are highlighted with lines. C: average amount of words (tokens) heard by a child or used by models. Note that for the audio dataset, many epochs of optimization over the same dataset take place.

guage algorithms remain remarkably inefficient, and thus calls for exploring novel neural architectures and training paradigms (Wolf et al., 2023; Hu et al., 2024). Part of this gap may be explained by non-language data pretraining. Indeed, Poli et al. (2024) showed that pretraining speech models on ambient sounds improves data-efficiency to then develop phonemic representations. Similarly, Orhan et al. (2025) showed that pretraining on ambient sounds lead to the detection of algebraic auditory structures.

Modeling the developmental trajectory of representations. This emergence strengthens earlier work on toy models. For example, (Kemp & Tenenbaum, 2008) showed that hierarchical concept structures can sequentially emerge from feature spaces by iteratively searching over a structured hypothesis space. More formally, Saxe et al. (2014; 2019) showed mathematically that hierarchical representations emerge sequentially in deep linear networks during training. Although we did not observe a sequential emergence of semantic categories (see Fig. A in appendix), we observe a sequential emergence across linguistic structures. Finally, Lavechin et al. (2025); Chen et al. (2024) show that self-supervised learning from raw speech and text yields a staged emergence of phonetic, early lexical structure and syntax without explicit labels. The present work complements these approaches by testing a unified code within self-supervised neural networks for the different linguistic structures posited by linguists.

Limitations Our findings come with several limitations. First, the developmental ordering between syntax and lexical semantics, a central question in cognitive science (Gleitman, 1990; Pinker, 1994), remains unresolved. Second, we report aggregate scores and do not stratify performance by simple versus complex syntactic phenomena or by token/phoneme frequency. Third, our evaluation relies exclusively on structural probing of internal representations, rather than on prediction-based tests at the phonemic, syntactic, or lexical levels (Schatz et al., 2013; Warstadt et al., 2020; Hill et al., 2015; Nguyen et al., 2020). Thus, our conclusions concern representation rather than inference and may not fully characterize downstream performance. Fourth, optimizing a probe to match pairwise distances might nevertheless be too restrictive. This approach might nevertheless be too restrictive. For example, the distance between "dog" and "river" does not have to be as precisely captured as the distance between "dog" and "cat". To show that our approach can be generalized beyond distance learning, while simultaneously strengthening our results, we present supplementary results using a "contrastive" probe, inspired by (Nickel & Kiela, 2017), which optimizes the extraction of the topology of the structure, irrespective of the distance. On measures of topological fidelity (unlabeled undirected attachment score for syntax, rank for syntax and semantic), this probe achieved better scores than the distance probe, while revealing similar emergence dynamics (Fig. J in appendix). These results prove that our approach can be used with different probing objectives.

Impact and Future Directions The present generalization of Hewitt & Manning (2019)'s structural probe to lexical semantics and phonology provides a unifying framework to characterize the structure and emergence of linguistic representations in neural networks. This approach reduces the notion of 'representation' to (squared) Euclidean distances between elements in specific subspaces of the activation space. This approach also provides a mechanistic explanation: As linear readout is a core operation of any neuron, linear subspaces make these language representations readily accessible to any subsequent steps (Kriegeskorte, 2015; Higgins et al., 2018; Park et al., 2024), including attention mechanisms. Our results, like previous interpretability works, suggest the existence of laws governing the geometry of neural networks' activations (Elhage et al., 2022; Hernandez et al., 2024; Jha et al., 2025; Park et al., 2025; Landauer & Dumais, 1997; Huh et al., 2024). Perhaps most excitingly, this work delineates a path to model the brain bases of language acquisition (Di Liberto et al., 2023; Nakagi et al., 2025; Evanson et al., 2025).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning and Computational linguistic. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020a. Curran Associates Inc. ISBN 9781713829546.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

Barrault, L., Chung, Y.-A., Meglioli, M. C., Dale, D., Dong, N., Duquenne, P.-A., Elsahar, H., Gong, H., Heffernan, K., Hoffman, J., Klaiber, C., Li, P., Licht, D., Maillard, J., Rakotoarison, A., Sadagopan, K. R., Wenzek, G., Ye, E., Akula, B., Chen, P.-J., El Hachem, N., Ellis, B., Gonzalez, G. M., Haaheim, J., Hansanti, P., Howes, R., Huang, B., Hwang, M.-J., Inaguma, H., Jain, S., Kalbassi, E., Kallet, A., Kulikov, I., Lam, J., Li, D., Ma, X., Mavlyutov, R., Peloquin, B., Ramadan, M., Ramakrishnan, A., Sun, A., Tran, K., Tran, T., Tufanov, I., Vogeti, V., Wood, C., Yang, Y., Yu, B., Andrews, P., Balioglu, C., Costa-jussà, M. R., undefinedelebi, O., Elbayad, M., Gao, C., Guzmán, F., Kao, J., Lee, A., Mourachko, A., Pino, J., Popuri, S., Ropers, C., Saleem, S., Schwenk, H., Tomasello, P., Wang, C., Wang, J., and Wang, S. Joint speech and text machine translation for up to 100 languages. *Nature*, 637 (8046):587–593, January 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08359-z. URL http://dx.doi.org/10.1038/s41586-024-08359-z.

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and Van Der Wal, O. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Boudreault, P. and Mayberry, R. I. Grammatical processing in american sign language: Age of first-language acquisition effects in relation to syntactic structure. *Language and Cognitive Processes*, 21(5): 608–635, August 2006. ISSN 1464-0732. doi: 10.1080/01690960500139363. URL http://dx.doi.org/10.1080/01690960500139363.

Chen, A., Shwartz-Ziv, R., Cho, K., Leavitt, M. L., and Saphra, N. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=MO5PiKHELW.

Christophe, A., Millotte, S., Bernal, S., and Lidz, J. Bootstrapping lexical and syntactic acquisition. *Language and Speech*, 51(1–2):61–75, March 2008. ISSN 1756-6053. doi: 10.1177/00238309080510010501. URL http://dx.doi.org/10.1177/00238309080510010501.

Cohen, R., Geva, M., Berant, J., and Globerson, A. Crawling The Internal Knowledge-Base of Language Models. In Vlachos, A. and Augenstein, I. (eds.), *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1856–1869, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.139. URL https://aclanthology.org/2023.findings-eacl.139/.

De Heer Kloots, M. and Zuidema, W. Human-like linguistic biases in neural speech models: Phonetic categorization and phonotactic constraints in wav2vec2.0. In *Interspeech 2024*, pp. 4593–4597. ISCA, 2024. doi: 10.21437/Interspeech.2024-2490. URL https://www.isca-archive.org/interspeech_2024/deheerkloots24_interspeech.html.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, June 2021. doi: 10.1162/coli_a_00402. URL https://aclanthology.org/2021.cl-2.11/.

Di Liberto, G. M., Attaheri, A., Cantisani, G., Reilly, R. B., Ní Choisdealbha, Á., Rocha, S., Brusini, P., and Goswami, U. Emergence of the cortical encoding of phonetic features in the first year of life. *Nature communications*, 14 (1):7789, 2023.

Diego-Simón, P., D' Ascoli, S., Chemla, E., Lakretz, Y., and King, J.-R. A polar coordinate system represents syntax in large language models. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 105375–105396. Curran Associates, Inc.,

2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/be36e50757bf9cd280aa74f89a7d1c23-Paper-Conference.pdf.

Diego Simon, P. J., d'Ascoli, S., Chemla, E., Lakretz, Y., and King, J.-R. A polar coordinate system represents syntax in large language models. *Advances in Neural Information Processing Systems*, 37:105375–105396, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/be36e50757bf9cd280aa74f89a7d1c23-Abstract-Conference.html.

Dubois, J., Poupon, C., Thirion, B., Simonnet, H., Kulikova, S., Leroy, F., Hertz-Pannier, L., and Dehaene-Lambertz, G. Exploring the early organization and maturation of linguistic pathways in the human infant brain. *Cerebral Cortex*, 26(5):2283–2298, 2016.

Dupoux, E. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59, 2018. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2017.11.008. URL https://www.sciencedirect.com/science/article/pii/S0010027717303013.

Eilers, R. E., Gavin, W., and Wilson, W. R. Linguistic experience and phonemic perception in infancy: A crosslinguistic study. *Child Development*, 50(1):14, March 1979. ISSN 0009-3920. doi: 10.2307/1129035. URL http://dx.doi.org/10.2307/1129035.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy Models of Superposition, September 2022. URL http://arxiv.org/abs/2209.10652. arXiv:2209.10652 [cs].

Evanson, L., Lakretz, Y., and King, J. R. Language acquisition: do children and language models follow similar learning stages? In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12205–12218, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.773. URL https://aclanthology.org/2023.findings-acl.773/.

Evanson, L., Bulteau, C., Chipaux, M., Dorfmüller, G., Ferrand-Sorbets, S., Raffo, E., Rosenberg, S., Bourdillon, P., and King, J.-R. Emergence of language in the developing brain. Preprint, AI at Meta & Fondation Adolphe de Rothschild Hospital. Available at: https://ai.meta.com/research/publications/emergence-of-language-in-the-developing-brain/, May 12 2025.

Frank, M. C. Computational models of early language acquisition. *Current Opinion in Neurobiology*, 21(3): 381–386, 2011.

Frank, M. C. Large language models as models of human cognition. *PsyarXiv*, July 2023. doi: 10.31234/osf.io/wxt69. URL http://dx.doi.org/10.31234/osf.io/wxt69.

Friedmann, N. and Rusou, D. Critical period for first language: the crucial role of language input during the first year of life. *Current Opinion in Neurobiology*, 35:27–34, December 2015. ISSN 0959-4388. doi: 10.1016/j.conb.2015.06.003. URL http://dx.doi.org/10.1016/j.conb.2015.06.003.

Friedmann, N., Belletti, A., and Rizzi, L. Growing trees: The acquisition of the left periphery. *Glossa: a journal of general linguistics*, 6(1), 2021.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H. L., and Paul, T. D. Mapping the Early Language Environment Using All-Day Recordings and Automated Analysis. *American Journal of Speech-Language Pathology*, 26(2):248–265, May 2017. ISSN 1058-0360, 1558-9110. doi: 10.1044/2016_AJSLP-15-0169. URL http://pubs.asha.org/doi/10.1044/2016_AJSLP-15-0169.

Gleitman, L. The structural sources of verb meanings. *Language Acquisition*, 1(1):3–55, January 1990. ISSN 1532-7817. doi: 10.1207/s15327817la0101_2. URL http://dx.doi.org/10.1207/s15327817la0101_2.

Gomez, R. L. and Gerken, L. Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, 70(2):109–135, March 1999. ISSN 0010-0277. doi: 10.1016/s0010-0277(99)00003-7. URL http://dx.doi.org/10.1016/S0010-0277(99)00003-7.

Grünewald, S., Friedrich, A., and Kuhn, J. Applying occam's razor to transformer-based dependency parsing: What works, what doesn't, and what is really necessary. In Oepen, S., Sagae, K., Tsarfaty, R., Bouma, G., Seddah, D., and Zeman, D. (eds.), *Proceedings of the 17th International Conference on Parsing Technologies and the*

IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021), pp. 131–144, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.iwpt-1.13. URL https://aclanthology.org/2021.iwpt-1.13/.

Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., and Bau, D. Linearity of Relation Decoding in Transformer Language Models, February 2024. URL http://arxiv.org/abs/2308.09124. arXiv:2308.09124 [cs].

Hershcovich, D., Abend, O., and Rappoport, A. Universal Dependency parsing with a general transition-based DAG parser. In Zeman, D. and Hajič, J. (eds.), *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 103–112, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-2010. URL https://aclanthology.org/K18-2010/.

Hewitt, J. and Manning, C. D. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North*, pp. 4129–4138, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL http://aclweb.org/anthology/N19-1419.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Hill, F., Reichart, R., and Korhonen, A. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, December 2015. doi: 10.1162/COLI_a_00237. URL https://aclanthology.org/J15-4004/.

Hu, M. Y., Mueller, A., Ross, C., Williams, A., Linzen, T., Zhuang, C., Cotterell, R., Choshen, L., Warstadt, A., and Wilcox, E. G. Findings of the second babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. *arXiv preprint arXiv:2412.05149*, 2024.

Huh, M., Cheung, B., Wang, T., and Isola, P. The Platonic Representation Hypothesis, July 2024. URL http://arxiv.org/abs/2405.07987. arXiv:2405.07987 [cs].

Jawahar, G., Sagot, B., and Seddah, D. What does BERT learn about the structure of language? In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1356. URL https://aclanthology.org/P19-1356/.

Jha, R., Zhang, C., Shmatikov, V., and Morris, J. X. Harnessing the Universal Geometry of Embeddings, May 2025. URL http://arxiv.org/abs/2505.12540. arXiv:2505.12540 [cs] version: 1.

Ji, H., Patel, T., and Scharenborg, O. Predicting within and across language phoneme recognition performance of self-supervised learning speech pre-trained models, June 2022. URL http://arxiv.org/abs/2206.12489. arXiv:2206.12489 [eess].

Jusczyk, P., Friederici, A., Wessels, J., Svenkerud, V., and Jusczyk, A. Infants´ sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, 32(3):402–420, June 1993. ISSN 0749-596X. doi: 10.1006/jmla.1993.1022. URL http://dx.doi.org/10.1006/jmla.1993.1022.

Kachergis, G., Marchman, V. A., and Frank, M. C. Toward a "standard model" of early language learning. *Current Directions in Psychological Science*, 31(1):20–27, 2022.

Kemp, C. and Tenenbaum, J. B. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, August 2008. doi: 10.1073/pnas.0802631105. URL https://www.pnas.org/doi/10.1073/pnas.0802631105. Publisher: Proceedings of the National Academy of Sciences.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kohl, P. K. Early linguistic experience and phonetic perception: implications for theories of developmental speech perception. *Journal of Phonetics*, 21(1–2):125–139, January 1993. ISSN 0095-4470. doi: 10.1016/s0095-4470(19)31326-9. URL http://dx.doi.org/10.1016/S0095-4470(19)31326-9.

Korchinski, D. J., Karkada, D., Bahri, Y., and Wyart, M. On the emergence of linear analogies in word embeddings. *arXiv preprint arXiv:2505.18651*, 2025.

Kriegeskorte, N. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1(1):417–446, November 2015. ISSN 2374-4650. doi: 10.1146/annurev-vision-082114-035447. URL http://dx.doi.org/10.1146/annurev-vision-082114-035447.

Kuhl, P. K. Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, 5(11):831–843, November 2004. ISSN 1471-0048. doi: 10.1038/nrn1533. URL https://www.nature.com/articles/nrn1533. Publisher: Nature Publishing Group.

Kuhl, P. K. Brain mechanisms in early language acquisition. *Neuron*, 67(5):713–727, 2010.

Landauer, T. K. and Dumais, S. T. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

Lavechin, M., De Seyssel, M., Titeux, H., Bredin, H., Wisniewski, G., Cristia, A., and Dupoux, E. Can statistical learning bootstrap early language acquisition? a modeling investigation. *PsyArXiv preprint psyarxiv/rx94d/*, 2022. URL https://osf.io/preprints/psyarxiv/rx94d/. Publisher: PsyArXiv.

Lavechin, M., de Seyssel, M., Titeux, H., Wisniewski, G., Bredin, H., Cristia, A., and Dupoux, E. Simulating early phonetic and word learning without linguistic categories. *Developmental Science*, 28(2), January 2025. ISSN 1467-7687. doi: 10.1111/desc.13606. URL http://dx.doi.org/10.1111/desc.13606.

Limisiewicz, T. and Mareček, D. Introducing orthogonal constraint in structural probes. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 428–442, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.36. URL https://aclanthology.org/2021.acl-long.36/.

Martin, K. C., Ketchabaw, W. T., and Turkeltaub, P. E. Plasticity of the language system in children and adults. *Handbook of clinical neurology*, 184:397–414, 2022.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech 2017*, pp. 498–502, 2017. doi: 10.21437/Interspeech.2017-1386.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., and Levin, L. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In Matsumoto, Y. and Prasad, R. (eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3475–3484, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1328/.

Müller-Eberstein, M., van der Goot, R., and Plank, B. Probing for labeled dependency trees. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7711–7726, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.532. URL https://aclanthology.org/2022.acl-long.532/.

Nakagi, Y., Tada, K., Yoshino, S., Nishimoto, S., and Takagi, Y. Triple phase transitions: Understanding the learning dynamics of large language models from a neuroscience perspective. *arXiv preprint arXiv:2502.20779*, 2025.

Nguyen, T. A., Seyssel, M. d., Rozé, P., Rivière, M., Kharitonov, E., Baevski, A., Dunbar, E., and Dupoux, E. The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling, 2020. URL http://arxiv.org/abs/2011.11588.

Nickel, M. and Kiela, D. Poincar\'e embeddings for learning hierarchical representations, 2017. URL https://arxiv.org/abs/1705.08039v2.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. Universal Dependencies v1: A multilingual treebank collection. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S. (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1659–1666, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1262/.

Orhan, P., Boubenec, Y., and King, J.-R. The detection of algebraic auditory structures emerges with self-supervised learning. *PLOS Computational Biology*, 21(9):1–23, September 2025. doi: 10.1371/journal.pcbi.1013271. URL https://doi.org/10.1371/journal.pcbi.1013271. Publisher: Public Library of Science.

Parcollet, T., Nguyen, H., Evain, S., Zanon Boito, M., Pupier, A., Mdhaffar, S., Le, H., Alisamir, S., Tomashenko, N., Dinarelli, M., Zhang, S., Allauzen, A., Coavoux, M., Estève, Y., Rouvier, M., Goulian, J., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., and Besacier, L. Lebenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of french speech. *Computer Speech & Language*, 86:101622, 2024. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2024.101622. URL https://www.sciencedirect.com/science/article/pii/S0885230824000056.

Park, K., Choe, Y. J., and Veitch, V. The Linear Representation Hypothesis and the Geometry of Large Language Models, July 2024. URL http://arxiv.org/abs/2311.03658. arXiv:2311.03658 [cs].

Park, K., Choe, Y. J., Jiang, Y., and Veitch, V. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bVTM2QKYuA.

Pasad, A., Chou, J.-C., and Livescu, K. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 914–921, 2021. doi: 10.1109/ASRU51503.2021.9688093.

Pasad, A., Chien, C.-M., Settle, S., and Livescu, K. What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics*, 12:372–391, 04 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00656. URL https://doi.org/10.1162/tacl_a_00656.

Pinker, S. How could a child use verb syntax to learn verb semantics ? In *The Acquisition of the Lexicon*. The MIT Press, 10 1994. ISBN 9780262286688. doi: 10.7551/mitpress/1075.003.0017. URL https://doi.org/10.7551/mitpress/1075.003.0017.

Pinker, S. *Language Learnability and Language Development*. Harvard University Press, Cambridge, MA and London, England, 1996. ISBN 9780674042179. doi: doi:10.4159/9780674042179. URL https://doi.org/10.4159/9780674042179.

Poli, M., Schatz, T., Dupoux, E., and Lavechin, M. Modeling the initial state of early phonetic learning in infants. *Language Development Research*, 5(1), 2024.

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ.

Rousso, R., Cohen, E., Keshet, J., and Chodroff, E. Tradition or Innovation: A Comparison of Modern ASR Methods for Forced Alignment, June 2024. URL http://arxiv.org/abs/2406.19363. arXiv:2406.19363 [eess] version: 1.

Sainburg, T., Mai, A., and Gentner, T. Q. Long-range sequential dependencies precede complex syntactic production in language acquisition. *Proceedings of the Royal Society B*, 289(1970):20212657, 2022.

Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, 2014. URL http://arxiv.org/abs/1312.6120.

Saxe, A. M., McClelland, J. L., and Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, June 2019. doi: 10.1073/pnas.1820226116. URL https://www.pnas.org/doi/10.1073/pnas.1820226116. Publisher: Proceedings of the National Academy of Sciences.

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., and Dupoux, E. Evaluating speech features with the minimal-pair ABX task: analysis of the classical MFC/PLP pipeline. In *Interspeech 2013*, pp. 1781–1785. ISCA, 2013. doi: 10.21437/Interspeech.2013-441. URL https://www.isca-archive.org/interspeech_2013/schatz13_interspeech.html.

Seyssel, M. d., Lavechin, M., Adi, Y., Dupoux, E., and Wisniewski, G. Probing phoneme, language and speaker information in unsupervised speech representations. In *Interspeech 2022*, pp. 1402–1406, 2022. doi: 10.21437/Interspeech.2022-373. URL http://arxiv.org/abs/2203.16193.

Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., and Manning, C. D. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014.

Tenney, I., Das, D., and Pavlick, E. BERT rediscovers the classical NLP pipeline. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Tesnière, L. *Eléments de Syntaxe Structurale*. Klincksieck, Paris, 1959.

Vong, W. K., Wang, W., Orhan, A. E., and Lake, B. M. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511, 2024.

Waldis, A., Perlitz, Y., Choshen, L., Hou, Y., and Gurevych, I. Holmes \ensuremath\recorder A Benchmark to Assess the Linguistic Competence of Language Models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647, 2024. doi: 10.1162/tacl_a_00718. URL https://aclanthology.org/

2024.tacl-1.88/. Place: Cambridge, MA Publisher: MIT Press.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., and Bowman, S. R. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. doi: 10.1162/tacl_a_00321. URL https://aclanthology.org/2020.tacl-1.25/.

Wolf, L., Tuckute, G., Kotar, K., Hosseini, E., Regev, T., Wilcox, E., and Warstadt, A. Whisbert: Multimodal text-audio language modeling on 100m words. *arXiv preprint arXiv:2312.02931*, 2023.

Yuan, Y., Jiang, Y., and Tu, K. Bidirectional transition-based dependency parsing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7434–7441, July 2019. ISSN 2159-5399. doi: 10.1609/aaai.v33i01.33017434. URL http://dx.doi.org/10.1609/aaai.v33i01.33017434.

Zada, Z., Nastase, S. A., Aubrey, B., Jalon, I., Michelmann, S., Wang, H., Hasenfratz, L., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Devore, S., Flinker, A., Devinsky, O., Goldstein, A., and Hasson, U. The "Podcast" ECoG dataset for modeling neural activity during natural language comprehension. *Scientific Data*, 12 (1):1135, July 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-05462-2. URL https://www.nature.com/articles/s41597-025-05462-2. Publisher: Nature Publishing Group.

A. Appendix

Table A. Datasets used to train syntactic, lexical semantic and phonological probes. Annotations of linguistic structures can be manual or automatic, marked by M and A respectively.

Name	Modality	Annotation
UD–EWT	text	UD (M)
WordNet–Nouns	text	WordNet (M)
UD–EWT–TTS	speech	UD (M) + Phonemes (A)
WordNet–Nouns–TTS	speech	WordNet (M)

Table B. Parameters of the linear probe. All probes are trained with the AMSGrad variant (Reddi et al., 2018) of the Adam optimizer (Kingma & Ba, 2014) For each combination of (model, task), we perform a grid search across learning rate, spanning linearly 20 values from 0.0000001 to 0.01, and report the best learning rate. Unlike its sensitivity to learning rate, the probe is more robust to batch size and probe initialization value. The probe is randomly initialized from a uniform distribution between -0.00001 and 0.00001. For each model, all hyperparameters are constant across layers. When the model size is not given, it means that the same hyperparameters were used across all model sizes.

Structure	**Model**	**Learning rate**	**Dim**	**Epochs**	**Probe init**	**batch size**
phonemic	Wav2vec 2.0	0.00001	768	100	0.00001	1 set of 200 phonemes
semantic	Wav2vec 2.0 30-94M	0.0000075	200	200	0.00001	300 set of 12 words
semantic	Wav2vec 2.0 315M	0.0005	200	200	0.00001	300 set of 12 words
semantic	pythia	0.0000025	200	300	0.00001	300 set of 12 words
semantic	Llama 2.0	0.00005	200	300	0.00001	300 set of 12 words
syntactic	Wav2vec 2.0	0.00001	200	2	0.00001	300 sentences
syntactic	pythia	0.000025	500	1	0.00001	150 sentences
syntactic	Llama 2.0	0.00001	200	1	0.00001	300 sentences

Syntax Dataset

Our generation pipeline will be made available in the form of the ALS (Auditory Linguistic Structure) python repository upon acceptance. This pipeline takes as inputs an annotated text corpus (ConLLu format), and synthesizes the equivalent auditory dataset, along with a precise temporal alignment. We make available a cluster-friendly version of this pipeline, which allows the reproduction of the dataset in less than one hour.

Speech synthesis For each sentence in the English-Web-Treebank (EWT) Universal Dependencies Dataset, we synthesize a voice clip using the "seamless-m4t-v2-large" deep learning model. First, we remove from each sentence any remaining contractions using the contractions python package. Removing contractions allows the Force Aligner to separate in time contracted words, (like You're) that would be otherwise recognized as a single event. Second, for each sentence, the text is used as input to a "seamless-m4t-v2-large" model, with default English speaker voice. We save the generated sounds as a WAV file with a 16000 Hz sampling rate.

Alignment Forced Alignment (FA) is a necessary step to know when each word in a sentence occurs in its corresponding synthesized voice clip. FA was conducted using the Montreal Forced Aligner (MFA) model, employing the `english_us_arpa` corpus and dictionary. To quantify the effect of different aligners, we also reproduced our analyses with the NeMo Forced Aligner. This aligner produced a worse alignment when manually checked, in accordance with Rousso et al. (2024) which compared MFA with other aligners and found MFA to outperform. Indeed, everything else fixed, the syntactic and semantic scores to the NeMo-aligned corpus were poorer, which indicates measurement error induced by the alignment. Despite the great care we took for this forced alignment, all of our metrics remain a lower bound estimate of the true model's abilities, as there will necessarily be some alignment errors in the process.

The Montreal Forced Aligner does not map exactly each word in its input text to each annotated word in the input corpus. Some composite words can indeed be split by MFA. To avoid silent misalignment mistakes, we run a simple iterative merging algorithm to group set of split words. This procedure leaves us with a beginning and an end time for every annotated word in the corpus.

Table C. Set of audio and text models along with their training dataset, the model size, total number of training steps, and availability of checkpoints. Control audio models were taken from the following paper: (Orhan et al., 2025). All speech models repository starting with type are in the folder: NDEM/model-wav2vec2

Training Dataset	Model Size	Training Steps	Checkpoints	Hugging face repository
Audioset	30M	100k steps	✓	_type-tiny_data-audiosetfilter_version-0
Audioset	94M	100k steps	✓	_type-base_data-audiosetfilter_version-2
FMA	30M	100k steps	✓	_type-tiny_data-fma_version-0
FMA	94M	100k steps	✓	_type-base_data-fma_version-2
LeBench	315M	200k steps	x	LeBenchmark/wav2vec2-FR-7K-large
Librispeech	30M	100k steps	✓	_type-tiny_data-librispeech_version-0
Librispeech	94M	100k steps	✓	_type-base_data-librispeech_version-4
Librispeech	94M	400k steps	✓	_type-base_data-librispeech_version-2
Librispeech	94M	400k steps	x	facebook/wav2vec2-base
Librispeech	315M	200k steps	x	facebook/wav2vec2-large
The Pile	70M-1.4B	143k steps	✓	EleutherAI/pythia-Xb
Meta Internal corpus	7B	500k steps	x	meta-llama/Llama-2-7b
Meta Internal corpus	13B	500k steps	x	meta-llama/Llama-2-13b

Table D. Scores of the semantic probe, (max pearson scores across layers)

Model	Training data	Semantic score (pearson correlation)
Wav2vec 2.0 30M	environmental sound	0.044
Wav2vec 2.0 94M	environmental sound	0.065
Wav2vec 2.0 30M	music	0.080
Wav2vec 2.0 94M	music	0.070
Wav2vec 2.0 315M	French	0.126
Wav2vec 2.0 30M	English	0.08
Wav2vec 2.0 94M	English	0.104
Wav2vec 2.0 315M	English	0.147
Pythia 70M	The Pile	0.118
Pythia 160M	The Pile	0.166
Pythia 410M	The Pile	0.22
Pythia 1B	The Pile	0.27
Pythia 1.4B	The Pile	0.29
Llama 2 7B	Meta Internal Corpus	0.42
Llama 2 13B	Meta Internal Corpus	0.415

Finally, we correct the information of the input corpus (ConLLu file), in the case of contractions. Decontracted sets of words are kept split as several elements (this should be done already in the ConLLu), while for re-contracted word (as can't becoming cannot), we keep only the first ConLLu token of the set of tokens associated with the contractions.

Semantic Dataset

Semantic category In the context of WordNet, we defined a semantic category as the set composed of a synset and all synsets that can reach this synset through hypernymy relationship (Fig A in appendix). For example, the noun graph root is "entity.n.01", and therefore all elements in the graph are part of the category "entity.n.01". Similarly, all synsets below "mammal.n.01" are part of the "mammal" category, but "animal.n.01" which is a hypernym of "mammal.n.01" is not.

Dataset generation The mapping between WordNet synsets and the set of nouns is neither surjective nor injective. Words are often polysemic, such that preserving words associated to a single synset would restrict our investigation to a small subgraph of the WordNet graph. Conversely, synsets are associated with multiple sets of words (lemmas). To deal with these two constraint, we can nevertheless take advantage of the fact that some synsets are over-precise and characterize a

very particular use of the words. For example, time is associated to 15 synsets, among which: time.n.05: 'the continuum of experience in which events pass from the future through the present to the past', and time.n.02: 'an indefinite period (usually marked by specific attributes or activities)'.

Consequently, to generate a dataset of pairs (synset,lemma) from the WordNet graph, we applied the following heuristic. (function wordnet_unisemic_wordlist in the ALS package). First, we collected all synsets in the WordNet noun graph. Each synset is associated with a list of lemmas (each lemma is a word or multiple words). We then repeated the following procedure: Given this list of synset, we first, whenever it was possible, assigned to a synset all lemmas for which no other synset has this lemma in its list. We then removed from the list of lemmas associated to this synset all lemmas that were in the list of other synsets. This allowed to repeat the procedure, as some lemmas then became assigned to a unique synset. In short, this first procedure selects for all synsets the lemma that is most unique to this synset.

This procedure still left a large number of polysemic lemmas, each associated with multiple synsets. Among those synsets, there are leaf synsets which have no hyponyms and are therefore very precise sense. We consequently pruned these leaf synsets out (removing a total of 9593 leafs synsets). The rationale was that the non-contextual evaluation of the lexical semantics of a polysemic word should rather be on its general sense than its precise and context-specific sense. Finally, we preserved in the graph the synset whose remaining list had at least one lemma associated with only one synset. Given this subset of synsets and their associated list of unique lemmas, we ultimately selected the lemma that had the highest word frequency according to the wordfreq package.

Cross-validation folds for Wordnet Next, we divided this large graph into train and test sets, using a splitting strategy that ensured that for each semantic category (defined as a synset and all its hyponyms), 20% of its members were in the test set. To be able to quantify how well each semantic category was coded in the model, we had to put in the test folds a subset of the words for each category. This assignment was not trivial, and we consequently came up with a simulated annealing procedure to obtain it. This procedure minimized a loss function that measures for all categories, the percentage of its elements in the test set. To be precise, the cost is defined as: $\sum_j \sum_{i \in C_j} (\delta_{i \in T} - 0.2|C_j|)^2$ where $\delta_{i \in T}$ is 1 if the synset i is in the test set, and C_j is the set of element in the category j, and the first sum iterates across all categories. Running a simulated annealing procedure involves randomly changing an element to be part of the training or testing set. This change is kept if the cost decreases ($\Delta cost < 0$) or if we sample a random number smaller than $\exp(-(\Delta cost)/\tau)$. τ controls the temperature and is progressively cooled. We run this simulated annealing procedure for 1000000 iterations with a cooling rate of 0.999995.

Restriction to the words appearing in librispeech In this work, we extensively compare audio and text models. Consequently, we decided to restrict our evaluation to words that appeared at least once in the Librispeech dataset. In the case of synsets associated with multiple words, we preserved only synsets for which all words appeared in the Librispeech dataset. The final dataset is consequently taken from the output of the cross-validation procedure and filtered to preserve only nouns appearing in the Librispeech dataset.

This procedure provided a reasonably large sample of (synset,lemma) pairs: 35,352, while avoiding synsets with major polysemic issues.

Emergence of semantic category does not replicate findings in toy models

The sequential emergence of increasingly complex semantic abilities is a property that was initially predicted in toy models (Saxe et al., 2014; 2019). In essence, this toy model predicts that larger semantic classes should emerge before precise semantic classes. We adapted these semantic classes by defining semantic categories in the Wordnet hierarchy (Fig A in appendix). To confirm whether these findings extended to transformer models, we then measured whether semantic categories emerged in the order of their size. Surprisingly, we found that this was not the case: class size did not predict the time of emergence of the lexical class (Fig A in appendix). For example, the separation of the main mammal subclasses, like "primate", "marsupial", "rodent", "carnivore", did not precede the separation of carnivore subclasses like "dog" or "feline". On the other hand, this supplementary analysis showed that semantic classes indeed exist, and improve with model size (Fig A in appendix), which strengthens our conclusions on the existence of a semantic graph in language models.

English exposed model instantiate partial semantic better than control models

To demonstrate that an English exposed model instantiate partial semantics in a significantly better way than a control model, we measured how well they could classify each semantic category. Remarkably, across most (167, 89%) large semantic categories, the English-exposed model had a better categorization score than control models. We show a scatter plot comparing these scores for early (5) and late (10) layers (Fig C in the appendix panel A-B). This difference increased through the layers of the model, peaking in layer 10 (Fig C in the appendix, panel C). Here, we focus on large categories with more than 100 words to show that this is driven by major semantic categories (a total of 187 categories), but this was also true when we looked at all of the 35532 semantic categories in the dataset.

UMAP visualizations confirm evaluations of semantic scores

We plot two UMAP visualisations before and after projection with the 200-D probe (Fig. D in appendix). Before projection, the space is only roughly divided into large semantic categories, presenting many small clusters without a clear organisation. After projection, the space is well organised into continuous clusters mirroring the WordNet tree hierarchy. Consequently, this unsupervised visualisation demonstrates that the WordNet tree is only instantiated in a subspace of the model activity, which is found by our probe.

Additional evaluation metrics

We provide in Fig. H additional results of the same syntactic and semantic probe but evaluated with two different metrics from the literature. The Unlabeled Undirected Attachment Score (UAS) measures the percentage of correct syntactic edges recovered by the probe (Hewitt & Manning, 2019). The rank score divides for each node i all other nodes into a positive ($P(i)$) set and negative ($N(i)$) set, depending on whether they are connected to i or not in the graph (semantic or syntactic). The rank score of this node is then the average rank of all positive nodes when sorting all nodes by increasing distance from i (Nickel & Kiela, 2017). The final rank score of a probe is then the median of these ranks scores across all nodes (Nickel & Kiela, 2017).

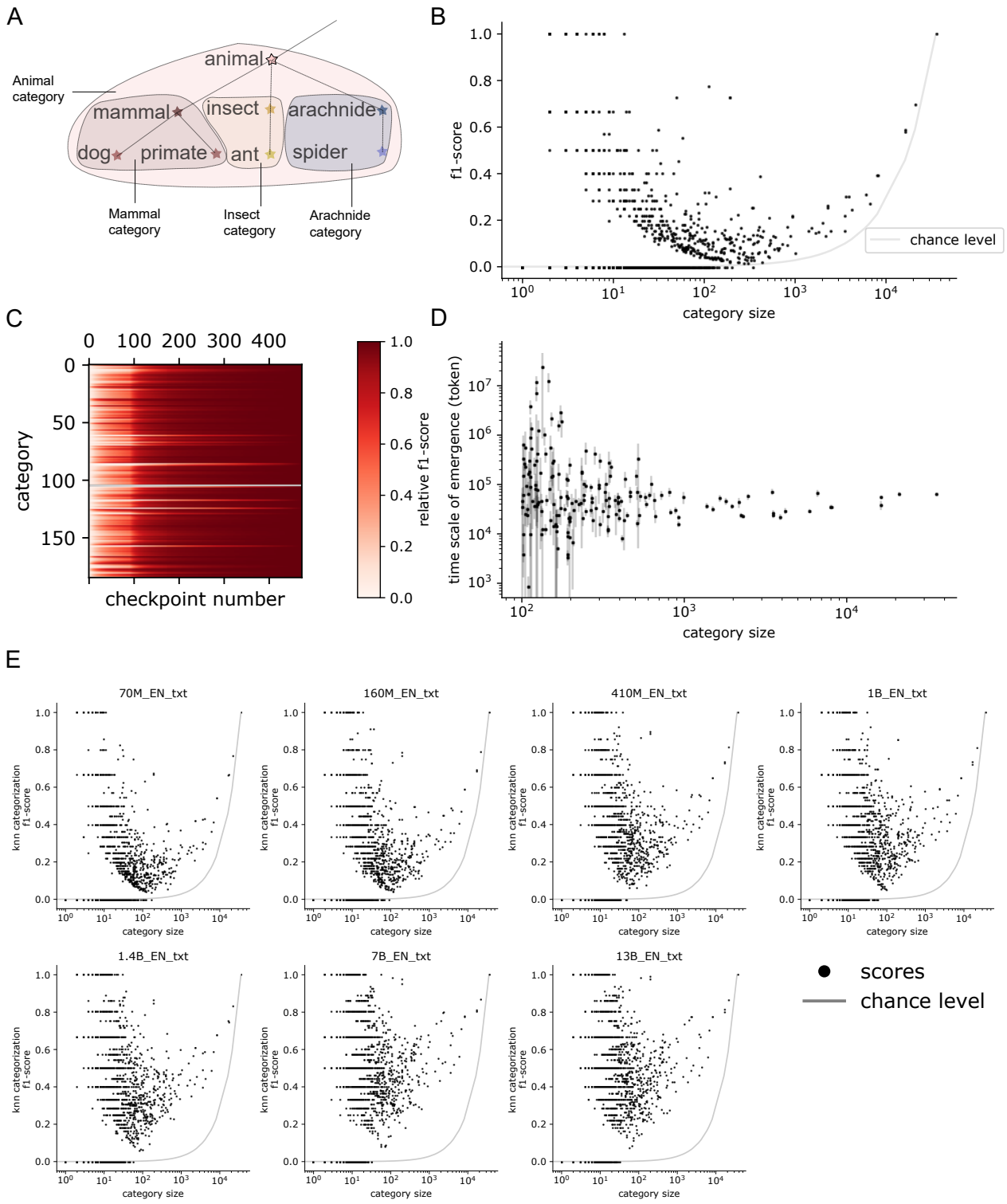

Figure A. A: Schematic of the definition of a category in a Wordnet hierarchy. B: f1-score of a 5-Nearest neighbor classification procedure, as a function of category size for a pretrained base Wav2vec 2.0 model. C: f1-score as a function of the checkpoint. D: Time scale of emergence as a function of the category size. Semantic categories do not form in the model in order of their size. E: f1-score for all language models in this study (Pythia suite from 70M to 1.4 billion parameters and Llama 2.0 large language models).

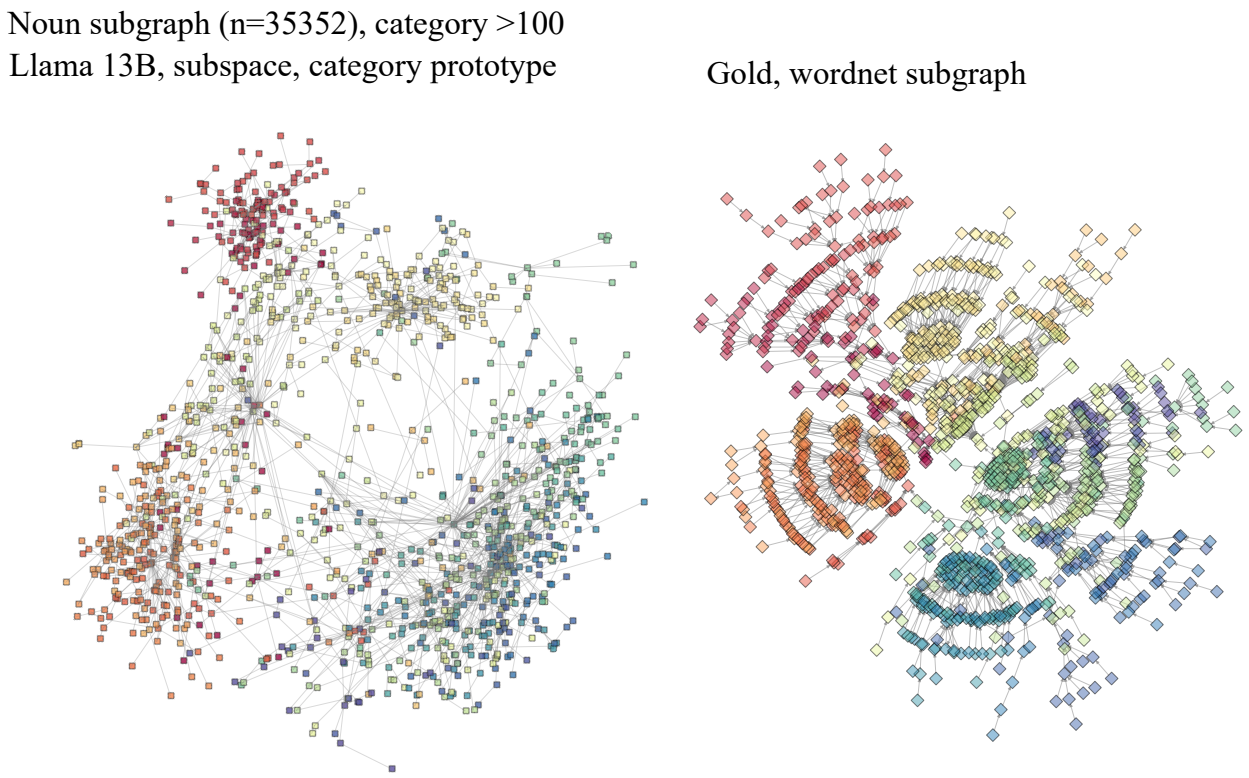

Figure B. Left: 2D projection of the noun subgraph probed from a Llama 2.0 models. Right: gold WordNet noun subgraph.

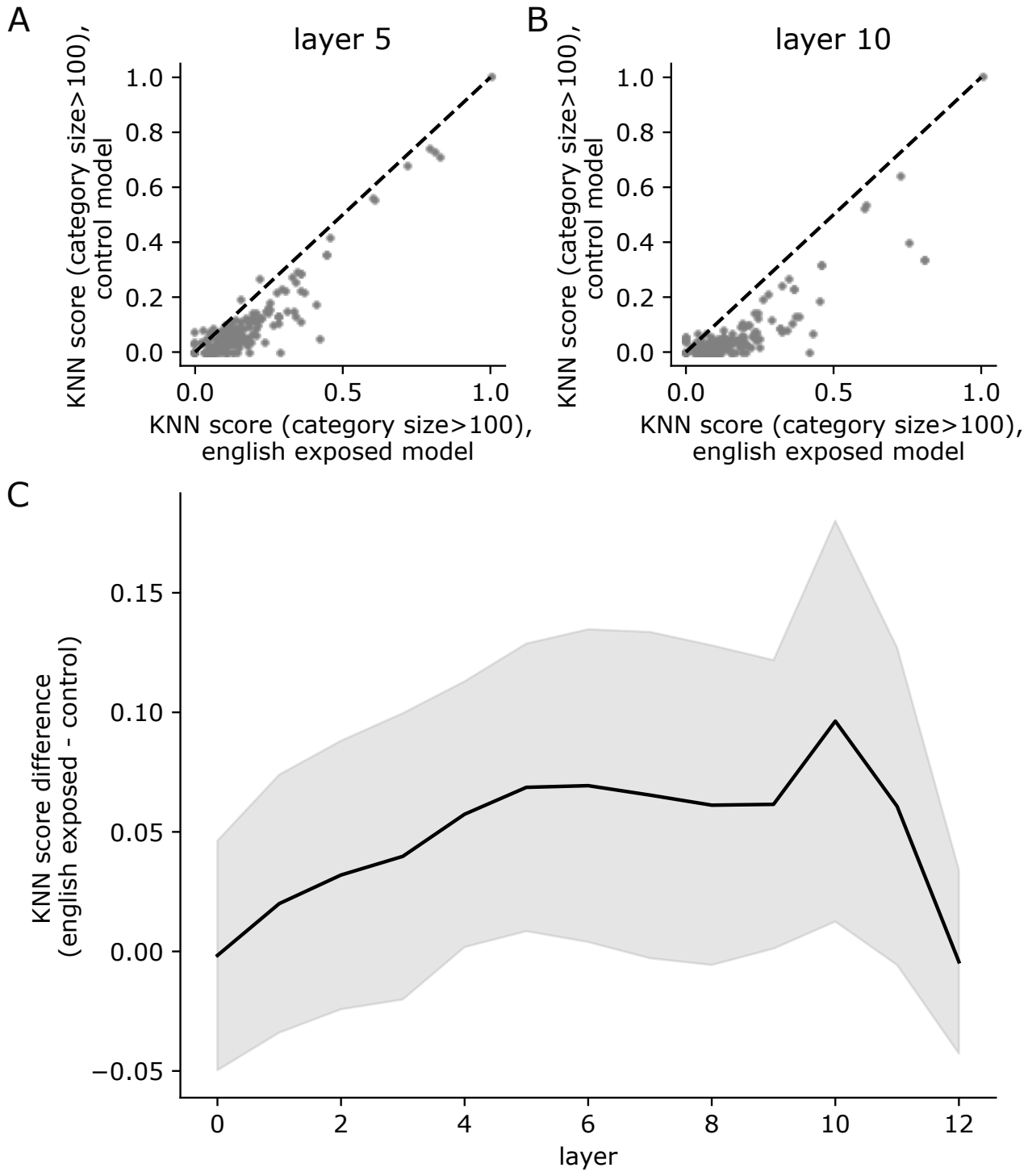

Figure C. A-B: Scatter plot of the f1-score of large semantic categories (more than 100 synsets) for an English-exposed Wav2vec 2.0 base model against a control model (Wav2vec 2.0 trained on environmental sounds) for layer 5 (A) and layer 10 (B). C: Difference of f1-score of the KNN classification between the English-exposed and control model, averaged across categories (black line) and with standard deviation (grey shadow) as a function of the layer.

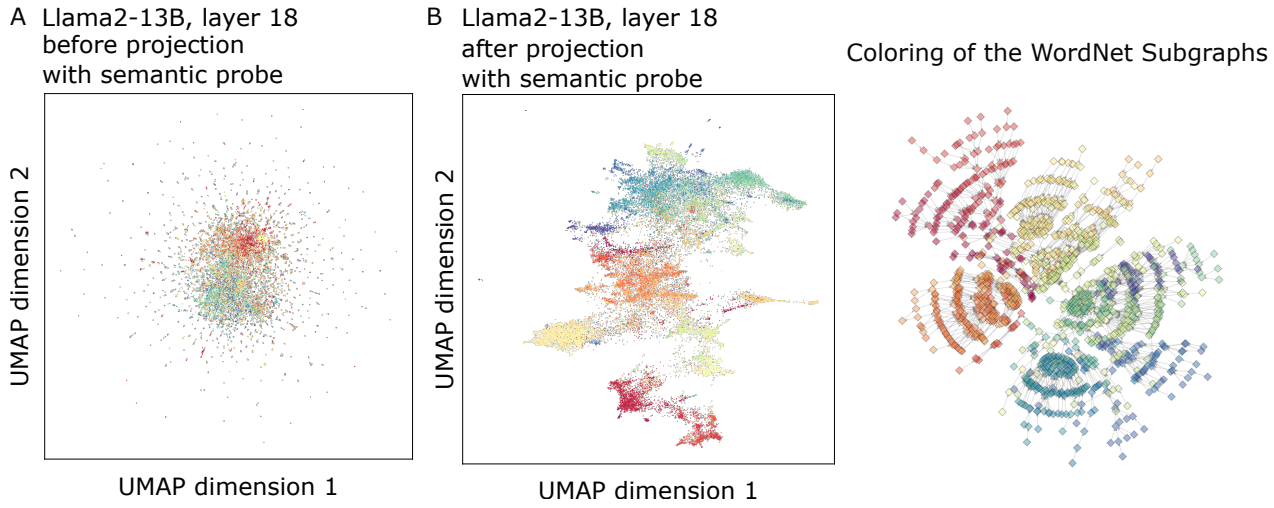

Figure D. A: UMAP visualization of the noun subgraph from the activity of a Llama 2.0 model. B: UMAP visualization after projection of a Llama 2.0 model with a 200-D semantic probe.

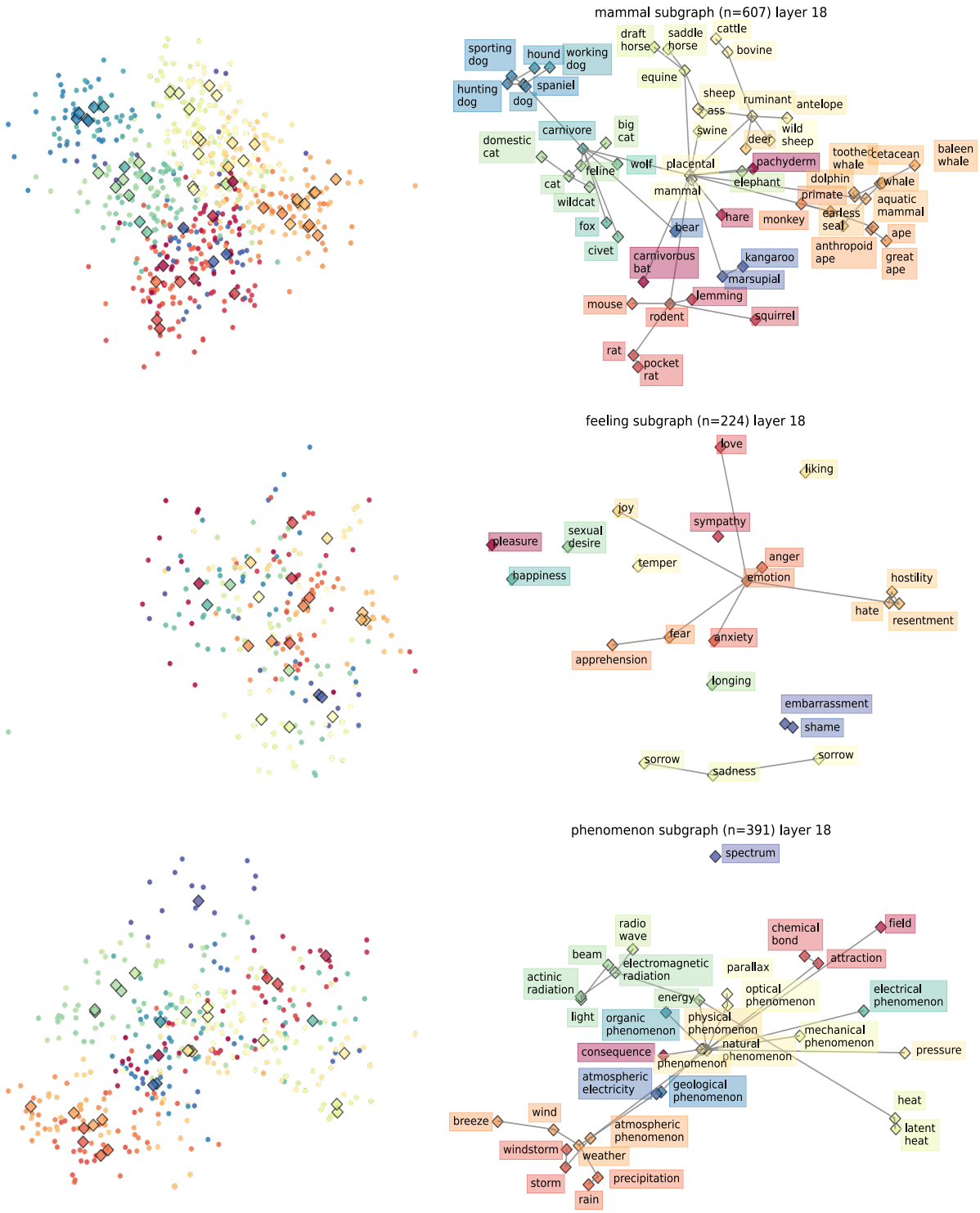

Figure E. Llama_13B 2D probing of mammal, feeling and phenomenon subgraphs of Wordnet

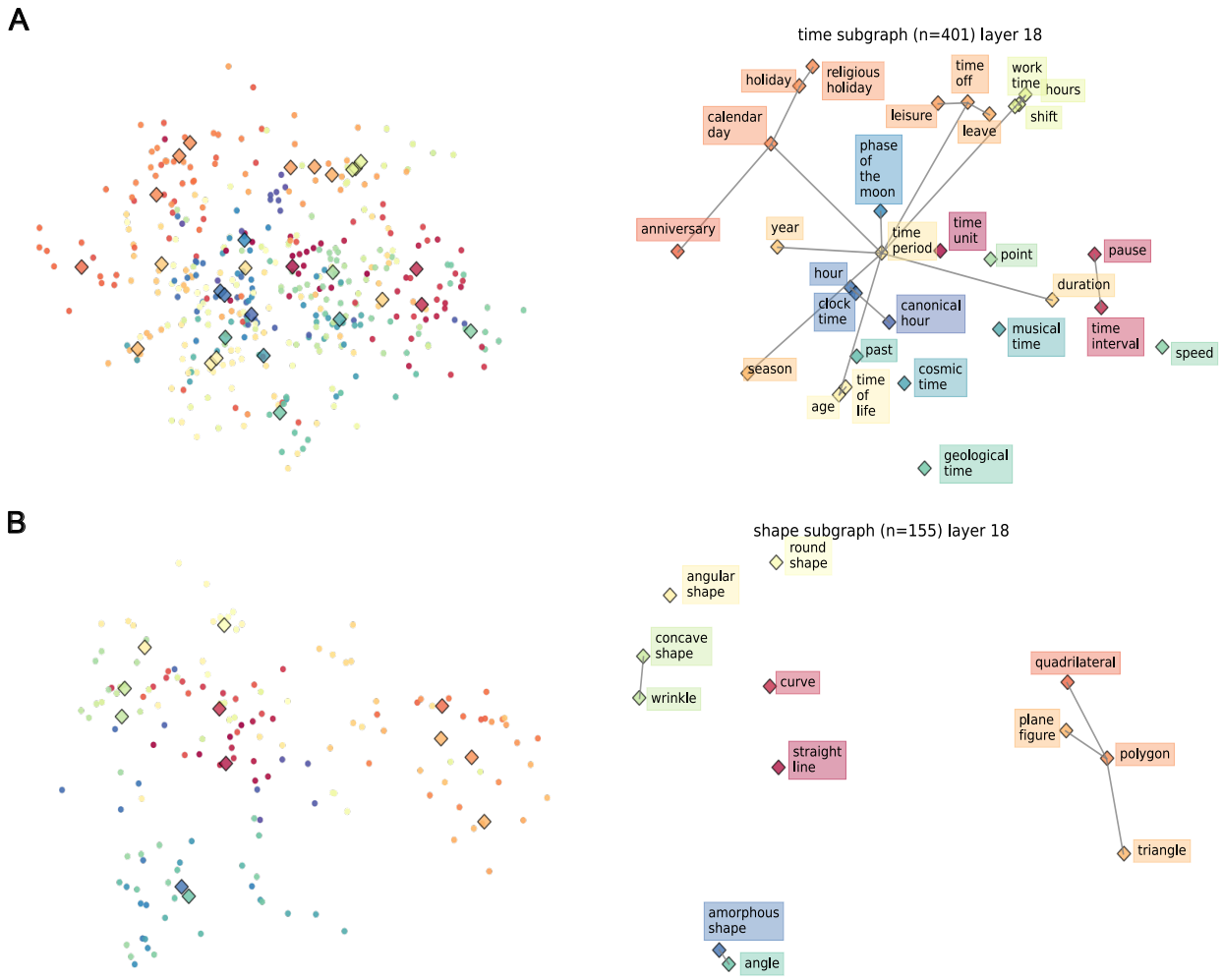

Figure F. Llama_13B 2D probing of time and shape subgraphs of Wordnet

Single unit analyses of semantic and syntactic subspaces

To gain some further insights on how single units interact in the coding of the semantic and syntactic subspace, we additionally ran 3 analyses on the syntactic and semantic probe of a Llama2 model.

We restricted these analyses to model layers for which the semantic and syntactic scores were above 80

First, to understand if two subspaces are aligned, we propose to define an alignment score as the mean square cosines of principal angles between two subspaces (Ake et al, 1973) (i,j): Second, we studied the norm of each unit's probe weights. Third, we quantified how well each unit's activity could be predicted by the semantic and syntactic subspace activity.

We first wanted to quantify whether semantic and syntactic subspaces recruited a similar pattern of units in the model. To do so, we computed the alignment between the left singular vector of each probe. More precisely, we first compute the left singular vector of each probe (over the model units). These vectors quantify how much each unit contributes to the activity of a particular subspace direction. Because they form an orthonormal basis from the same space, we can then compute an alignment score (Ake et al, 1973) as:

$$Alignment(B_i, B_j) = \frac{1}{N}||_i^T V_j||_F^2 = \frac{1}{N}\sum_{n=1}^{N} X_n = \frac{1}{N}\sum_{n=1}^{N} \cos^2(\theta_n) \quad (1)$$

The alignment score is 0 for orthogonal subspaces (all angles are 90°) and 1 for perfectly aligned subspaces (all angles are

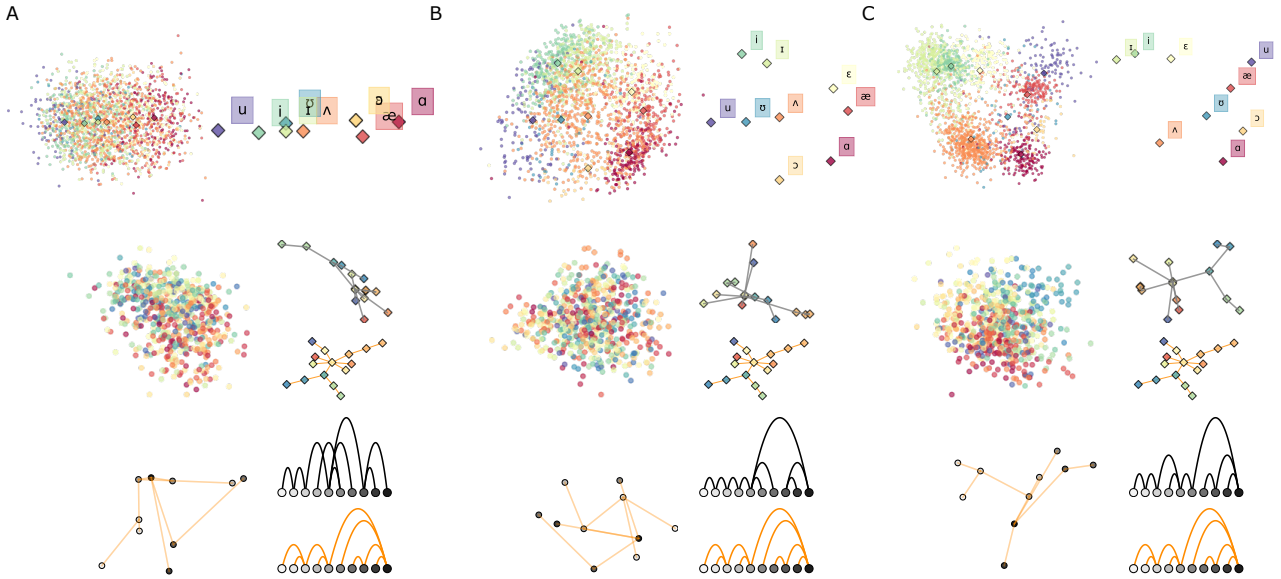

Figure G. A, B and C: 2d visualizations of the phonemic and semantic space along with one example syntactic structure for 3 pertaining steps.

0°). A rough interpretation of in-between scores can be made by taking their arc-cosine. Overall, scores below 0.6 are likely to indicate that the two spaces are not aligned.

We observed a small alignment between the semantic and syntactic subspace of Llama models (max: 0.042, min:0.038)(Fig. I, in appendix, panel A), with a relatively constant value across layers. These results indicate that semantic and syntactic subspaces are mostly unaligned and do not recruit a similar pattern of units in the model.

We next addressed the question of the role of each network unit. We measured for each model unit the norm of its probe vector. (Fig. I, in appendix, panel B) After plotting the distribution of these norms, we observed that probes weights were spread across all units of the layer. This is expected as, following previous work (Hewitt & Manning, 2019), we do not explicitly regularize the norm of the probe, which could have favored the discovery of a sparser solutio We also observed more outliers for the syntactic subspace than the semantic subspace. Some units (40 to 90, depending on the layer) had outlier norm, ranging from 2 to 3 times the median norm. Some units (max 20) were both outliers for the syntactic and semantic subspace. We then repeated these measurements across syntactic and semantic-coding layers of the model (Fig. I, in appendix, panel C), which revealed that an increasing number of units were both outliers for the syntactic and semantic subspace.

These results suggest that the representation of syntactic and semantic subspace is widely distributed across the network units, with a few sets of units being more strongly read out by the probe, although their contribution is mostly washed out by the influence of the majority. Consequently, these subspaces form distributed and distinct, although low-dimensional, patterns of activity in the initial activity space.

To better understand the activity of each individual units, and how we could understand them in terms of syntactic and semantic subspace, we then turned to an encoding approach Indeed, the previous analysis on the optimized probe is not entirely convincing, as the probe is not regularized, such that it could give a high norm to two units while making their contribution cancel out. To deal with that issue, we tried to predict the model unit response from the activity of the syntactic and semantic subspace. Indeed, if some units' activity can be predicted and explained by linear combination of the subspace activity, but not by simpler features, they are likely strongly participating in the syntactic or semantic code.

To remove contamination due to training and testing on particular datasets, we explore a novel set of sentences taken from the Podcast ECOG dataset (Zada et al., 2025). This dataset is essentially a transcription of a 30-minute long podcast composed of 5305 words. We gathered the activations of Llama2 by inputing block of 100 words from the podcast transcript in the model and gathering the average activity of all tokens associated with each individual words. We then used the semantic and syntactic probed previously estimates to project each layer activity into a semantic and syntactic subspace. For this

analysis, we only kept the 34 layers out of the 40 layers for which the semantic and syntactic scores were above 80% of their respective maximal scores across layers. This was done to make sure the subspace instantiated true semantic and syntactic structure rather than another projection of the model activity.

We then ran an independent ridge regression with 5 outer and 5 inner cross-validation folds (sequential fold block), from the subspace activity to each individual unit response. We then partitioned the variance of each regression into the variance explained by the semantic and syntactic subspace $R2 = R2_{semantic} + R2_{syntax}$. Remarkably, a small number of units (36 in layer 15) had variance strongly predicted by the activity of the syntactic subspace ($R2_{syntax} > 60\%$). Interestingly, these units matched with the outlier units found in the previous analysis (Fig. I, in appendix, panel D).

We then controlled this regression for the effect of known univariate syntactic features. For that purpose, we used the 96 univariate "syntactic features" available with the Podcast dataset, which are made as one-hot encoding of part-of-speech and dependency relationship categories. These features are supposed to encode syntactic properties while not capturing the distance of the syntactic tree. We ran two regressions, one with only the univariate features as predictors, and the second with the univariate features and the subspace activity as predictors. (Fig. I, in appendix, panel E) Remarkably, adding the subspace activity strongly increased the variance explained, indicating that syntactic features were not sufficient to explain these units' responses. Crucially, this was the case in outlier units, whose variance was consequently explained by the syntactic tree-coding subspace instead of univariate semantic features. We additionally verified by eye that these units had no outlier response, computing their mean response across each part-of-speech or dependency category; no clear explanation of their tuning could be found in this manner. In addition, these units' responses were not strongly correlated across the words of the dataset (mean correlation: 0.002, max across two units: 0.39, standard deviation: 0.125), stressing that they had distinct roles in the encoding of syntactic information. We repeated this analysis across all syntax-coding layers of the model (Fig. I, in appendix, panel F), finding a similar conclusion.

Unlike for syntax, no such effect was observed for the semantic subspace: this subspace of activity was not sufficient by itself to predict a large variance of any single unit response.

Together, these results suggest that the instantiation of semantic structure is a distributed process, spread across the units of the model at each layer. The instantiation of syntactic structure is also a distributed process, but with a few units specializing in syntax-coding.

Probing topology instead of distances

To complement our distance-based probe, we designed a novel contrastive probe that optimizes the extraction of the topology of the structure, irrespective of the distance. This probe is inspired by contrastive learning approaches of embedding on WordNet (Nickel & Kiela, 2017), and more generally by contrastive learning objectives used in word-embedding vectors.

Contrastive probe objective. The contrastive probe optimizes the contrastive objective of (Nickel & Kiela, 2017), where we replaced the embeddings with a linear projection from the model activity, and optimize the projection. More precisely, given a node i from a graph S, we define the set of positive nodes $P(i)$ as the nodes directly connected to this node. We define set of negative nodes $N(i)$ as all other nodes. We then optimize the following objective:

$$\hat{B} = argmin_{B \in \mathbb{R}^{k,p}} \sum_{i \in S} -\log \left(\frac{\sum_{j \in P(i)} \exp(-||(h_i - h_j)B||_2^2)}{\sum_{k \in N(i)} \exp(-||(h_i - h_k)B||_2^2)} \right) \tag{2}$$

In practice, to compute this objective efficiently, we use batched gradient descent, sampling a positive node and a set of negative node. All hyperparameters searches are performed identically to the distance-based probe, with the exception of the number of epochs, which was the same for syntactic probes, but increased to 1000 epochs for semantic probes. Indeed, we observed a slower convergence of the semantic probe with the contrastive objective in term of the number of epochs, but since the computational cost of each epoch is low, we could afford this increase in the number of optimization epochs. The contrastive probe is indeed more memory efficient than the distance-based probe, as it does not require to compute and store all pairwise distances between nodes.

Evaluations Unlike distance-based probe, the contrastive probe does not optimize the distances between all pairs of nodes. Consequently, it is expected that distance-based evaluations, like spearman correlation show lower scores for this probe.

Instead, it should reach higher score on topological evaluations, we use the Unlabeled Undirected Attachment Score (UUAS) for syntax, and rank-based evaluations for syntax and semantics.

Results Results obtained with the contrastive probe confirmed the results obtained with the distance-based probe. The topology of the semantic WordNet graph emerged in middle-late layers of both text and audio models (Fig. J in appendix, panel A). English-exposed model showed better instantiation of this topology than music-exposed, environmental sounds-exposed or French-exposed control models (Fig. J in appendix, panel B). And the instantiation of this topology increased with model size (Fig. J in appendix, panel B). As observed with the distance-based probe, instantiation of semantic topology progressively emerged during training (Fig. J in appendix, panel C), and saturated for the audio model. Similar results were obtained for the syntactic tree topology (Fig. J in appendix, panels D to I), where we evaluated the instantiation of the syntactic structure through UUAS and rank scores. Distance-based probe and contrastive probe led to similar qualitative conclusions although they offer difference quantitative scores because they optimize different objectives (Fig. K in appendix). The distance-based probe led to higher scores for distance-based evaluations (spearman correlation), while the contrastive probe led to higher scores for topological evaluations (UUAS and rank scores).

Together these results confirm the validity of our measurements using the distance-based probe.

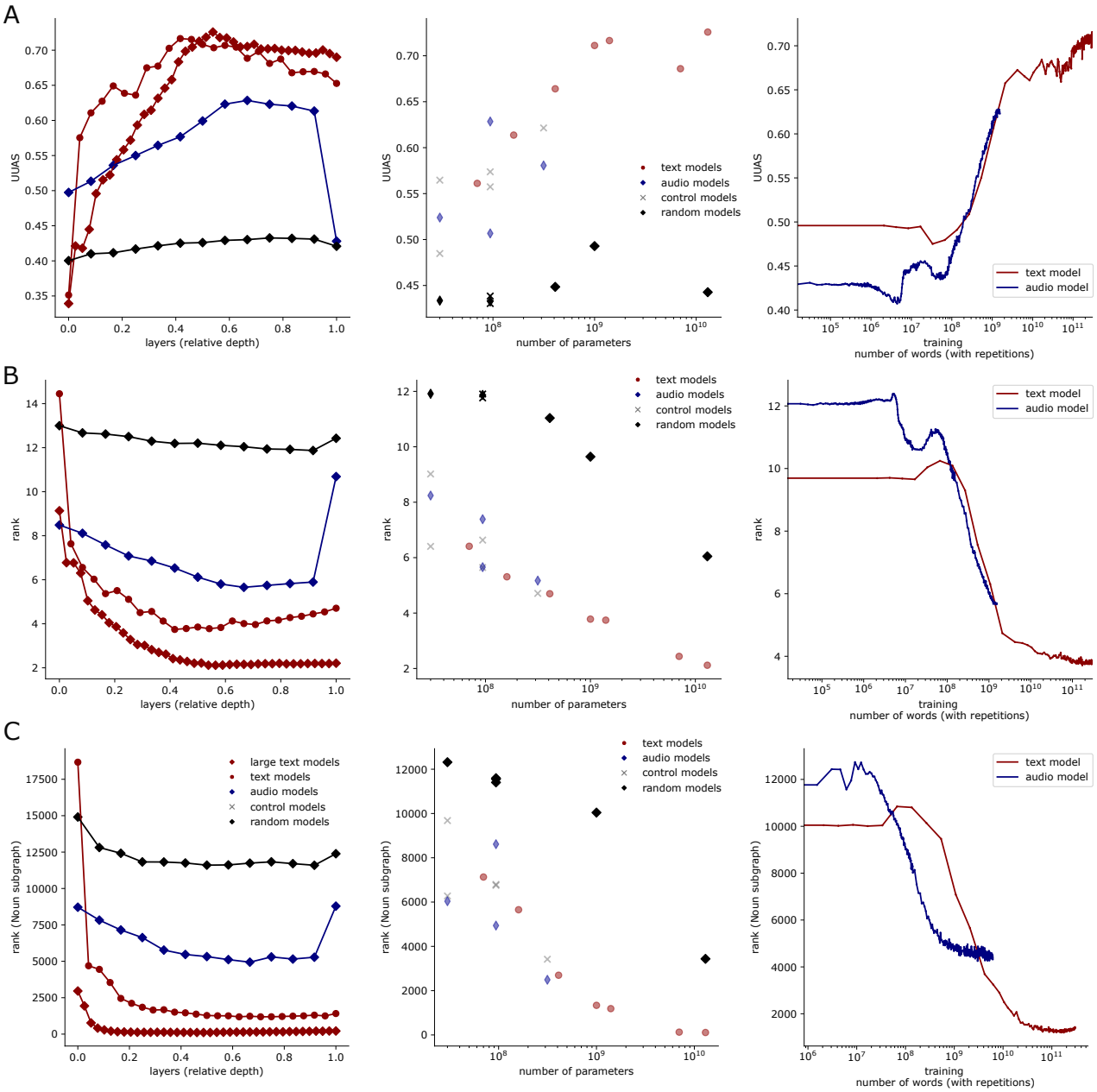

Figure H. A: Same as Fig.4, but evaluating the Unlabeled Undirected Attachment Score (UUAS) B. Same as Fig.4, but evaluating the rank score. C: Same as Fig.3, but evaluating the rank score

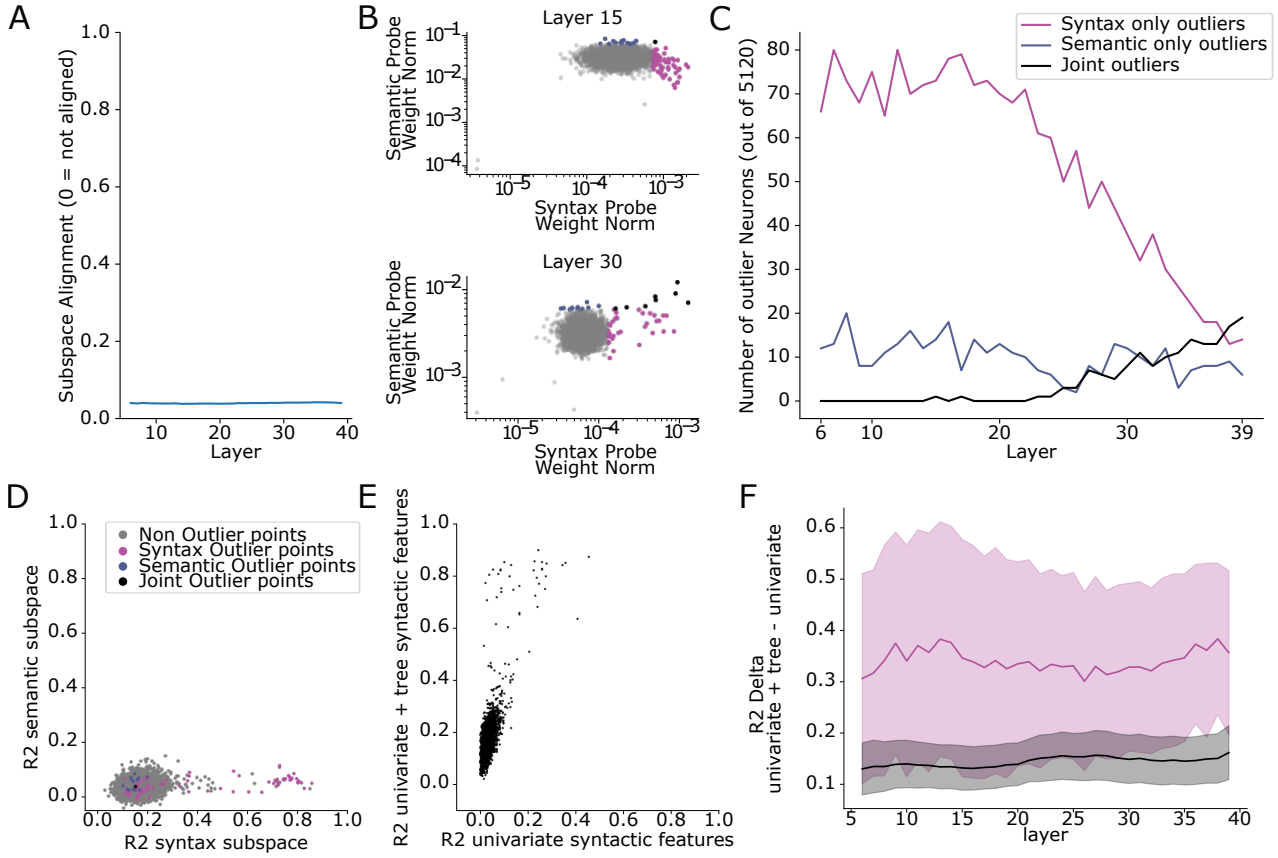

Figure I. Subspace alignment and single unit analyses. A: Subspace alignment between the syntactic and semantic subspace across layers instantiating syntax and semantic structure of a Llama2-13B model. B: Scatter plot of syntax and semantic probe norm for all units of layer 15 (top) and layer 30 (bottom) of a Llama2-13B model. Units with outlier probe weight norm are highlighted in purple (syntax), navy (semantic) and black (joint outliers). C: Number of outlier units for syntax, semantics, and joint outliers across layers instantiating syntax and semantic structure of a Llama2-13B model. D: Encoding R2 score of each unit of layer 15 of a Llama2-13B model from the syntactic subspace activity and semantic subspace activity (R2 partitioning). To show that outlier units of the structural probe were indeed explained by the syntactic subspace activity, these units' markers are colored in purple. E: Unique contribution of the syntactic subspace activity compared to simpler univariate syntactic features for each unit of layer 15 of a Llama2-13B model. F: Difference of R2 score (subspace + univariate features - univariate features) across layers instantiating syntax of a Llama2-13B model.

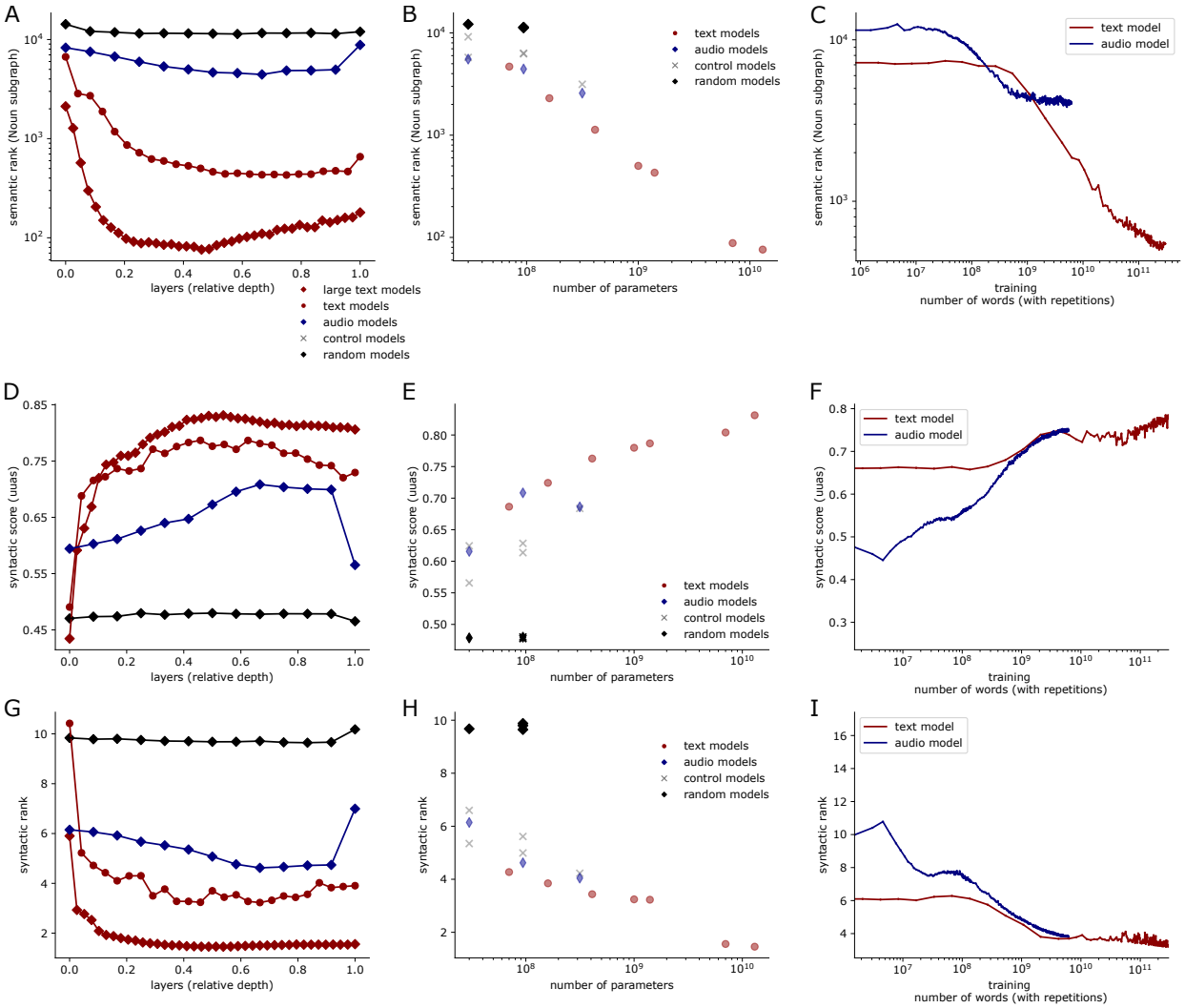

Figure J. A: Semantic rank score of a contrastive probe for a large text (Llama-13B), text (pythia-1.4B), audio (Wav2vec 2.0-94M base), and random (Wav2Vec 2.0-94M base) models. B: Semantic rank scores of a contrastive probe for all models as a function of the model size. C: Semantic rank score of a contrastive probe as a function of the quantity of pertaining for the text (pythia-1.4B) and audio (Wav2vec 2.0-94M base) models. D-F: Same as A-C but for the syntactic dataset evaluated with the UUAS score. G-I: Same as D-F but for the rank score.

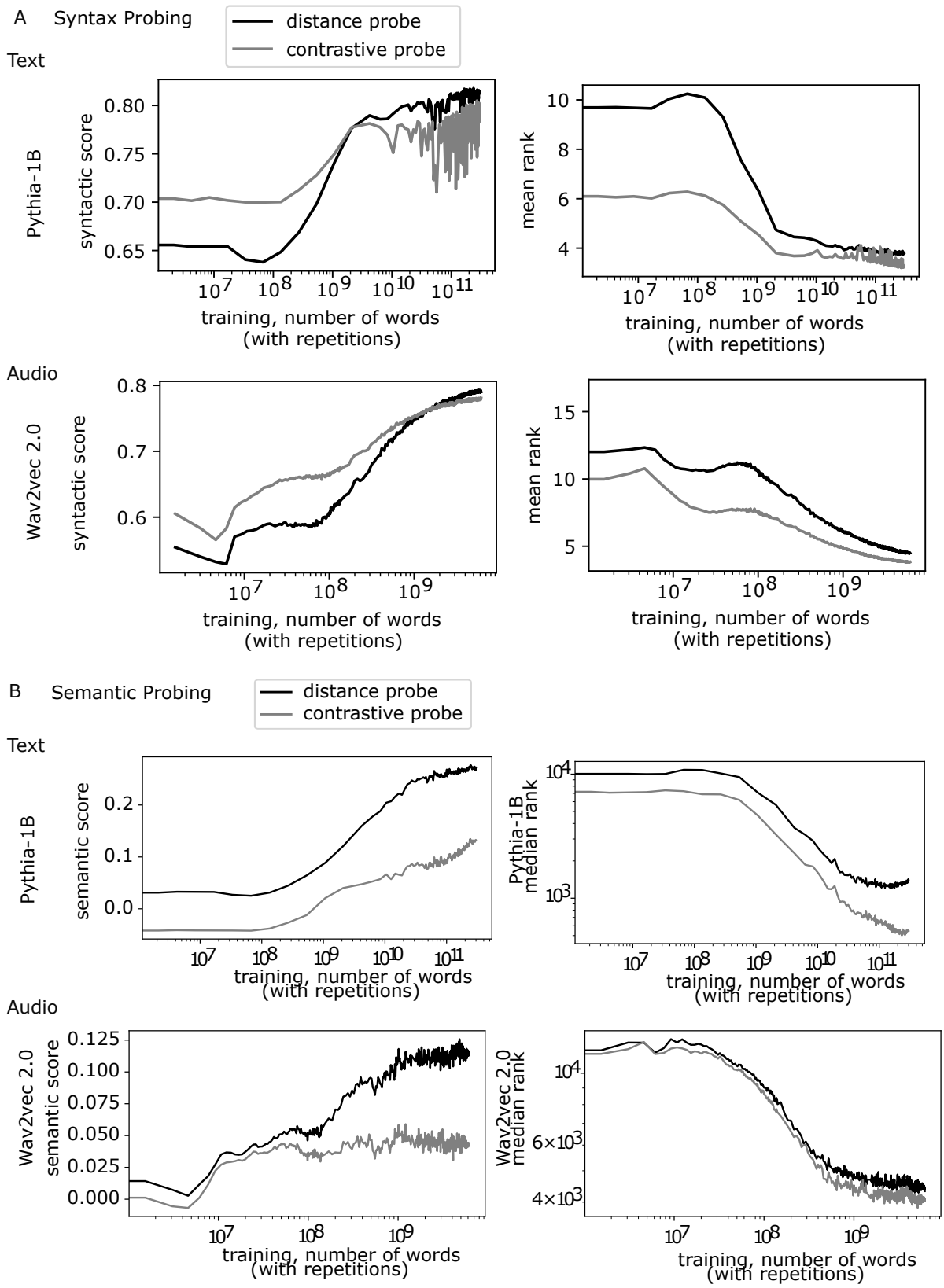

Figure K. A: Score of the structural (distance-based) and contrastive probe on the syntax dataset for a text (pythia-1.0B) and audio (Wav2vec 2.0-94M base) model. From left to right, we report for every checkpoint the spearman correlation (syntactic score) and mean rank score. B: Same as A, but for the semantic dataset, reporting spearman correlation (semantic score) and median rank score.