ON THE PROTO-ROLE PROPERTIES INFERRED BY TRANSFORMER LANGUAGE MODELS

MATTIA PROIETTI GIANLUCA E. LEBANI
ALESSANDRO LENCI

ABSTRACT: In recent years Language Models have taken the Computational Linguistics community by storm. Nevertheless, very little is known of the kind of linguistic knowledge that these systems are able to infer from the input they receive. In this work we address whether, and to what extent, different architectures of different sizes are able to encode the semantic content of Dowty (1989, 1991)'s semantic proto-roles in the contextual embeddings that they generate. Following Lebani & Lenci (2021) and Proietti *et al.* (2022), we test four different models by creating a linear mapping between the generated contextualized embeddings and a semantic space built on the basis of the proto-roles annotations collected by White *et al.* (2016). For each model, the embeddings generated by the learned mapping were tested against the manual annotation of a set of previously unseen verbs in context, as well as qualitatively investigated to test to what extent they are able to model the semantic properties of the agent of the verbs participating in the so-called causative alternation. All in all, our results not only extend to more Transformer Language Models previous findings showing that proto-roles information is available in distributional semantic models, but also show that larger models are not necessarily better at modeling proto-role properties, in line with recent psycholinguistic evidence.

KEYWORDS: thematic proto-roles, semantic roles, distributional semantic models, contextual word embeddings, argument alternations.

1. INTRODUCTION[1]

Notwithstanding the number of major breakthroughs that radically changed the Natural Language Processing and Computational Linguistics scenario in

[1] This research was partly funded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 – "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme. The paper is the result of a joint work by the three authors. For the specific purposes of Italian Academy, Mattia Proietti is responsible for Sections 3 and 4, Gianluca E. Lebani for Sections 2 and 5, and Alessandro Lenci for Sections 1 and 6.

the last decade, the mechanisms used to represent meaning have not changed much. All the different flavours of such techniques are rooted in the same theoretical account, that of the distributional semantics and the distributional hypothesis. Indeed, the various models leverage the common idea that meaning can be depicted by means of inferring multidimensional numerical vectors (commonly referred to as *embeddings*) from the context of the occurrence of a given target word (Lenci 2018; Lenci & Sahlgren 2023). By means of encoding statistical correlations extracted from large corpora, embeddings can implicitly represent important semantic features about words and phrases in a way that is both compact and machine readable. Still today, in the era of generative and conversational AI (e.g., ChatGPT), word and sentence embeddings are crucial tools for giving these models access to the meaning of human language (Lenci 2023).

One technique to generate embeddings that has gained great popularity in recent times, is to leverage a special kind of neural language models which are based on the Transformer architecture (Vaswani *et al.* 2017) and can yield contextual word representations, hence called *contextual embeddings*. This kind of vector can represent the same word in different ways when it occurs in different contexts, thus helping to form richer and less ambiguous semantic representations, overcoming the capabilities of previous methods relying on static, hence context-free, vectors (Mikolov *et al.* 2013a). Transformers are complex neural networks originally composed of two principal neural blocks, called respectively the *encoder* and the *decoder* (Cho *et al.* 2014; Sutskever *et al.* 2014), enriched with a mechanism called *attention* to generate such context-aware word representations. Thanks to that complex computational tool, a neural language model based on transformers can represent a word by considering the surrounding context, incorporating information from neighboring words into the vector of the target word. Various kinds of Transformers exist, which can be differentiated on the basis of their specific architectures, training objective functions and sizes. In this work we will focus on models belonging to two major architectural families, the *encoder-only* and *decoder-only* Transformers, which take their names on the specific Transformer block of which they make use and whose characteristics will be briefly outlined below and in Section 3.1. We may refer to such models as *Transformer Language Models* (TLMs) or simply *Language Models* (LMs) in the context of the present work.

The greatest advancements in performance we witnessed lately by virtue of such models and the contextual embeddings they yield are not always paired with a deeper understanding of the nature and the actual content of these dense numerical representations. Therefore, a persistent halo of partial mystery keeps surrounding them. In these representations, indeed, information is distributed

across a set of dimensions that cannot be associated with a specific label or sets of labels. Nor is the linguistic behaviour of these models in learning such representations well understood, a fact that in the last years drove a surge in the number of studies focusing on the nature of linguistic knowledge that is acquired by deep neural networks (e.g., Ettinger 2020; Rogers *et al.* 2020; Baroni 2022). To make the situation even worse, recent evidence suggests that larger Transformer-based language models yielding lower perplexity (that is, models yielding a higher linguistic accuracy that arguably results in an improvement in the performance of downstream applications) are less predictive of human reading times (Oh & Schuler 2022, 2023) and of eye tracking measurements reflecting lexical access and early semantic integration (de Varda & Marelli 2023), arguably producing less cognitively accurate semantic representations.

Probably the most widely used tasks developed to identify the kind of information encoded by (word, sentence, or even multi-modal) embeddings are the so-called *probing tasks* (Conneau *et al.* 2018; Vulić *et al.* 2020), that is, classification tasks that predict a given property. A major problem of this approach, however, is that it still provides only indirect evidence about the content of these embedded vectors (Schwartz & Mitchell 2019).

An alternative technique to test whether word embeddings encode a specific type of knowledge, which is also more interpretable, is to train a machine learning model to learn a mapping between uninterpretable embeddings and a space populated by vectors whose dimensions are human-interpretable features. This technique, initially proposed by Mikolov *et al.* (2013b) for a different purpose, has been used to test whether embeddings encode cognitive features (Făgărășan *et al.* 2015), brain-based semantic features (Utsumi 2020; Chersoni *et al.* 2021) or the semantic properties of the agent and patient thematic roles (Lebani & Lenci 2021; Proietti *et al.* 2022). With the partial exception of Proietti *et al.* (2022), all the mentioned works focus on embeddings that associate a single representation for each lexical element, irrespective of its intended meaning in the actual context of use (i.e., static embeddings).

The work presented in these pages inscribes itself in this literature both by adopting the cited mapping technique and by focusing on the same semantic information explored by Lebani & Lenci (2021) and Proietti *et al.* (2022), that are the entailment relations that, according to Dowty (1989, 1991)'s approach, form the semantic content of the *proto-agent* and *proto-patient* role, defined as a cluster of properties that an argument possesses by virtue of its role in the event described by a predicate. To the best of our knowledge, other relevant works that explored the possibility to use a distributional approach to model a Dowty-inspired representation include Lebani & Lenci (2018), in which the authors developed a model able to represent the thematic role prop-

erties activated by a subset of English verbs and Rudinger *et al.* (2018) and Stengel-Eskin *et al.* (2020, 2021), in which contextual information is used to parse text into a Dowty-inspired semantic representation.

As in Proietti *et al.* (2022), we focus on contextual embeddings, i.e., representations that keep track of the different contexts in which a word occurs, associating different vectors with the different uses of the same word (for review, see Ethayarajh 2019; Liu *et al.* 2020). We however extend this literature by contrasting the contextual embeddings generated by different language models, organized along the following two, already cited, dimensions:

- *model architecture*: encoder-only, bidirectional models vs. decoder-only, autoregressive language models. In the former architectures, contextual embeddings are built by scanning the whole context of appearance of a given word, while in the latter only the previously seen elements in the sentence are considered. Usually, bi-directional models are most used for encoding texts and extracting word or sentence embeddings from them while left-to-right unidirectional models are more apt at generating text.

- *model size*: (relatively) larger vs. (relatively) smaller models in terms of parameters and structure of the neural architecture (i.e. number of layers) in order to test the long-lasting idea that "the larger the model, the better the representation" that is implicitly assumed in the literature.

We tested the contextual representations generated by these models in two experiments. In a first experiment, we created a linear mapping between each vector space generated by our models and the proto-roles annotations collected by White *et al.* (2016), leaving out a set of test sentences that were later used to measure the correlation between the human-annotated scores and the vectors generated by each mapping-embedding pair. We then moved to a qualitative analysis of how the generated embeddings are able to model the semantic properties of the agent of the verbs participating in the so-called *causative-inchoative alternation*.

This paper is organized as follows. While in Section 2 we introduce the main theoretical concepts of our analysis, i.e. the notion of semantic proto-role and the causative-inchoative alternation, Section 3 is devoted to the description and justification of the design of our experiments, whose results are reported in Section 4. Finally, in Section 5 we discuss the results reported in the previous sections and draw a few considerations about the ability of the tested models to encode proto-role information in their verb representation.

2. SEMANTIC PROTO-ROLES

Semantic roles, also known as *case relation*, or *thematic roles/relations*, are labels assigned to arguments based on their function in the event or situation described by a predicate (Levin & Rappaport Hovav 2005: ch. 2). Examples include the AGENT role, representing the *"animate and volitional initiator or doer of an action"*, and PATIENT/THEME, representing the *"entity undergoing the action and somehow affected by it"* (Pustejovsky & Batiukova 2019: 29).

Traditionally, semantic roles are treated as linguistic primitives describing a natural class of arguments. Such an approach, however, faces many challenges, including a lack of consensus on their definition, granularity, and identification, as reviewed by Dowty (1991: 553-559). It is indeed not uncommon to find sentences in which one or more arguments are ambiguous between different roles, as it is the case for *John* in the sentence *John ran into the house*: is *John* a AGENT, because he initiates the movement, a THEME, because he moves, or, as suggested by Jackendoff (1972), both, thus violating the uniqueness assumption that is implicit in most theories of thematic roles?

Recent years have seen the emergence of a novel approach, inspired by the seminal work of Dowty (1989, 1991), according to which thematic roles are bundles of properties or entailments imposed by the predicate over its arguments. Some of these properties are *verb-specific*, while others are *linguistic*, more abstract, properties that are licensed by many verbs. Dowty (1991) identified the two clusters of linguistic properties that he labelled as the PROTO-AGENT and the PROTO-PATIENT, here described in (1) and (2):[2]

(1) Contributing properties for the Agent Proto-Role (Dowty 1991: 572)
   a. volitional involvement in the event or state
   b. sentience (and/or perception)
   c. causing an event or change of state in another participant
   d. movement (relative to the position of another participant)
   (e. exists independently of the event named by the verb)

(2) Contributing properties for the Patient Proto-Role (Dowty 1991: 572)
   a. undergoes change of state
   b. incremental theme
   c. causally affected by another participant
   d. stationary relative to movement of another participant
   (e. does not exist independently of the event, or not at all)

[2] Parenthesis in 1.e and 2.e are present in the original paper.

Dowty's view suggests that PROTO-AGENTS and PROTO-PATIENTS tend to be realized in active sentences as subjects and objects, respectively, organized akin to the prototypes described by Rosch & Mervis (1975). This prototype structure allows for flexibility, as an argument is not obligated to exhibit all the entailments of a given proto-role. The categorization as an AGENT or PATIENT depends on the number of PROTO-AGENT and PROTO-PATIENT entailments received from the predicate. For example, the subject and object of the verb *to build* possess respectively all the properties outlined from (1) and (2), classifying them as exemplars of an AGENT and of a PATIENT instances, respectively (Dowty 1991: 572). Conversely, subjects of psychological predicates like *to fear* appear to be less "agentive", lacking PROTO-AGENT volitionality and event-causing entailments (Dowty 1991: 573).

2.1 Collections of proto-role entailments

Dowty (1991)'s perspective has found support through both psycho-linguistic evidence (McRae *et al.* 1997; Ferretti *et al.* 2001; McRae *et al.* 2005; Kako 2006a,b; Hare *et al.* 2009) and contributions from the Computational Linguistics and Natural Language Processing field (Reisinger *et al.* 2015; Lebani & Lenci 2018, 2021; Proietti *et al.* 2022). Our work is closely related to the corpus-based verification of Dowty's theory conducted by Reisinger *et al.* (2015) and refined by White *et al.* (2016). These scholars, indeed, not only demonstrated the validity of the proto-role hypothesis on large scale corpus-based data, but also provided two publicly available datasets of proto-roles annotations that are nowadays released as part of the Universal Decompositional Semantics Dataset (White *et al.* 2020).[3]

Inspired by Kako (2006b), Reisinger *et al.* (2015) developed a crowdsourcing annotation task in which each annotator was presented with a PropBank (Palmer *et al.* 2005) sense in which an argument was highlighted. Annotators were tasked with judging the plausibility of a property of the highlighted argument using a 5-point Likert scale, addressing questions of the form *"How likely or unlikely is it that ARG is sentient?"* that were selected in order to describe twelve role properties selected from the role hierarchy proposed by Bonial *et al.* (2011): *instigated*; *volitional*; *awareness*; *sentient*; *moved*; *physical existed*; *existed before*; *existed during*; *existed after*; *changed possession*; *change of state*; *stationary*. These authors collected judgments for over 9,000 arguments of nearly 5,000 verb tokens, spanning 1,610 verb sense IDs.

White *et al.* (2016) further refined Reisinger *et al.* (2015)'s protocol, enhancing the inventory of annotated properties and incorporating redundant an-

[3] Available online at the URL: `http://decomp.net`.

ROLE PROPERTY	HOW LIKELY OR UNLIKELY IS IT THAT
`instigation`	ARG caused the PRED to happen?
`volition`	ARG chose to be involved in the PRED?
`awareness`	ARG was/were aware of being involved in the PRED?
`sentient`	ARG was/were sentient?
`change of location`	ARG changed location during the PRED?
`existed before`	ARG existed before the PRED began?
`existed during`	ARG existed during the PRED?
`existed after`	ARG existed after the PRED stopped?
`change of possession`	ARG changed possession during the PRED?
`change of state`	ARG was/were altered or somehow changed during or by the end of the PRED?
`was used`	ARG was/were used in carrying out the PRED?
`was for benefit`	PRED happened for the benefit of ARG?
`partitive`	Only a part or portion of ARG was involved in the PRED?
`change of state continuous`	The change in ARG happened throughout the PRED?

TABLE 1: PROPERTIES OF THE REVISED STRATEGY DESCRIBED BY WHITE *et al.* (2016).

notations. Table 1 provides a complete overview of the properties discussed in White *et al.* (2016) and the questions used to elicit them. This revised protocol was used to annotate the Universal Dependencies English Web Treebank (version 1.2: Silveira *et al.* 2014), thus covering a wider range of genres than PropBank and working on a treebank annotated according to the Universal Dependencies (UD, de Marneffe *et al.* 2021). The dataset resulting from this improved protocol consists of 206,018 annotations (198,002 involving a NP argument) for 957 verbs in 2,793 sentences, with 4,607 verbal tokens (4,600 of which had at least one NP argument) and a total of 7,144 verb-argument pairs (6,142 if we consider nominal arguments only).

2.2 Causative-inchoative alternation

As a theoretical tool, semantic proto-roles can be used to define the linguistic patterns exhibited by verbs that undergo argument alternations – i.e., those verbs capable of appearing in multiple syntactic contexts Levin (2015). A case in point is that of the so called causative-inchoative alternation (Levin 1993; Levin *et al.* 1995; Levin & Rappaport Hovav 2005), a linguistic phenomenon (see below Example (5)) commonly treated as a cue of the Unaccusative Hy-

pothesis (Perlmutter 1978; Burzio 1986). Following the Unaccusative Hypothesis, among the intransitive verbs, there is an underlying split between two sub-groups, the so-called *unergative* as opposed to the *unaccusative*, each one characterized by a different syntactic configuration. While the former can be described as verbs taking an external argument but not a direct internal one, the latter, on the opposite, lack of an external argument but have in turn an internal direct one (Perlmutter 1978; Levin *et al.* 1995). In other words, from a semantic roles perspective, whatever specific theory is taken into account, the subject of the *unergative* verbs is considered as a realization of an agent-oriented role, while the subject of the *unaccusative* is thought to bear a more patient-oriented role. Thus, while the sole argument of an *unergative* verb like the one in example (3) is considered to be a subject both on the surface and on a deeper level, the one in (4) is considered to be an underlying object.

(3) *John worked*

(4) *John died*

Some verbs inside the *unaccusative* sub-category can be further differentiated because they allow both transitive and intransitive constructions, thus giving rise to the so-called causative-inchoative alternation. These are often referred to as "change of state" and "change of position" verbs (Levin 1993). *Unaccusative* verbs, participating in the alternation, can appear in double contexts as shown in the following example, where the verb *to break* appears in a transitive frame in (5a) and in an intransitive one in (5b):

(5) a. *John broke the window.*
b. *The window broke.*

In a sense, this phenomenon is not only a clear diagnostic of the division among intransitive verbs and a probing for the Unaccusative Hypothesis, but also a valuable lens to inspect the behaviour of semantic roles and their relationship with the syntactic arguments through which they are realized.

We might further analyze the sentences in (5) by means of event decomposition and formalize them through the following formal representation:

(6) a. x CAUSE[BECOME[BROKEN(y)]]
b. BECOME[BROKEN(y)]

The formalization in (6) helps to clarify two aspects. First, the transitive version of a predicate like *to break* is composed of two logical operators, CAUSE and BECOME, which are related to two predicates taking respectively an external argument (x) and an internal one (y). The intransitive version, on its side, seems to be a reduction of the first, relying only on the BECOME operator, which holds a single internal argument (y). The second important thing to point out is that the second argument of the transitive predicate overlaps with the first argument of the intransitive one. In other words, while y is the object in (6a), it is a subject in (6b). From a syntactic perspective, this is in tune with the observation that subjects of intransitive unaccusative verbs are underlying objects. From a semantic point of view, according to Dowty's theory of thematic proto-roles, this implies that in frames like that in (6b) subjects tend to be rather a realisation of the PROTO-PATIENT role than the PROTO-AGENT, while the standard intuition claims that the first argument of a predicate, generally the subject, is more likely to be characterized by PROTO-AGENT entailments. In the following sections, we will leverage this theoretical account to test the possibility to model this phenomenon computationally, by harnessing the semantic representations of verb embeddings.

3. METHODS

The recent literature in computational linguistics offers a series of successful applications of the methods we adopted here, which will be described in the remainder of this section. Ideed, training a linear mapping between word embeddings and human-generated features has been proven to be a simple, yet useful, probing task to investigate to what extent it is possible to retrieve human-like knowledge from complex neural network models (Vulić *et al.* 2020). For example, Făgărășan *et al.* (2015) learn a linear mapping to generate representative features for novel unseen target words. In a more cognitively oriented perspective, Utsumi (2020) and Chersoni *et al.* (2021) aim at decoding the content of word embeddings by mapping them onto the set of brain-based semantics features by Binder *et al.* (2016). The same approach is adopted by Lebani & Lenci (2021) and Proietti *et al.* (2022) who learn a linear mapping between diverse word embeddings and the collection of human judgements about the Proto-Roles properties of English verbs by White *et al.* (2016), we have reviewed in Section 2.1.

The present work extends Proietti *et al.* (2022) on both the theoretical and the technical sides. Two main experiments have been conducted. In the first one, we trained a machine learning model on a linear mapping between the embedding spaces obtained from several Transformer language models and a

target entailment semantic space made up of human ratings. We conducted this experiment over two target spaces, corresponding respectively to the `nsubj` (active subject) argument properties and the `dobj` (direct object) ones, for each chosen model. For the second experiment, we trained a linear regression model based on each embedding space using solely transitive verbs annotated with proto-roles properties for the `nsubj` argument. We then used each model to predict the properties of a set of 100 unseen sentences, containing 50 verbs participating in the inchoative-causative alternation, examples of which are given below in Section 4.2. Differently from Proietti *et al.* (2022), we compare four language models of different sizes and architectures.

3.1 Language Models

In our experiments, we have compared language models that are representative of two major types of transformer architectures:

- *encoder-only, bidirectional models* – they are trained on a masked language modeling task inspired to the Cloze test: some tokens from the input are randomly masked (e.g., *The dog is* [MASK] *a red ball*), and the network training objective is to predict the original masked words based on their context. These language models are bidirectional, because a word is predicted by considering both the preceding and the following items;

- *decoder-only, autoregressive language models* – they are unidirectional and trained to predict the conditional probability of a target word (e.g., *ball*) given the preceding context (e.g., *The dog is chasing a red* ...). These language models are also called generative, because they are optimized to generate the most likely text sequences following a certain context.

Since the model size (i.e., the number of the network parameters) is crucial in determining the model's performance (Kaplan *et al.* 2020), for each architectural type, we selected a large model and a much smaller one, to investigate how this factor affects their ability to encode Proto-Role information.

- **BERT** (Devlin *et al.* 2019): The large version of the BERT transformer, a pre-trained encoder language model with an architecture made of 24 layers, a vector size of 1024 and a total of 336M parameters.[4]

[4] `https://huggingface.co/bert-large-uncased`.

- **DistilBERT** (Sanh *et al.* 2019): This is a smaller version of the original BERT model, obtained through a process called distillation. It results in having ∼40% less parameters than the `bert-base-uncased` model, reducing the total to ∼60M parameters. The performances of this little model are often comparable with those of the original one.[5]

- **GPT2-XL** (Radford *et al.* 2019): The largest version of the original open-source GPT2 architecture by OpenAI, an autoregressive decoder-only model with a total of 1.5b parameters and a vector size of 1600.[6]

- **Pythia70m** (Biderman *et al.* 2023): Pythia70m belongs to a suite of 8 models based on the GPT2 architecture and built in a scaled fashion, ranging from 70m parameters to 12b parameters, for the purpose of doing interpretability research over different model sizes. This is the smallest of our set of selected models, with 6 layers, 8 attention heads and a dimensionality of 512.[7]

3.2 Data and Rating-based semantic spaces

Starting from the dataset in White *et al.* (2016), built upon sentences extracted from the EWT corpus (Silveira *et al.* 2014), we derived two different rating spaces for the `nsubj` and for the `dobj` arguments, containing representations of target verbs tokens. Similarly to Lebani & Lenci (2021) and Proietti *et al.* (2022) the rating spaces have 14 dimensions, each of which corresponds to one of the properties elicited in White *et al.* (2016) and shown above in Table 1. The ratings for each token have been averaged across different annotators and the resulting aggregated token representation received an index corresponding to the sentence *id* found in the EWT corpus, to keep track of different instances of the same verb type. In this way, we ended up in having indexes of verbs of the form *verb.id* where *id* is the identifier of the sentence in which a given verb has been found. The corpus of raw sentences has been reconstructed by tracking the sentence *id* found in the White *et al.* (2016) dataset to retrieve these target sentences directly from the English Web Treebank. Due to tokenization alignment problems between the models' tokenizers and the original split found in the EWT corpus, we filtered out a group of sentences (∼100) obtaining spaces of slightly different sizes than those found in Proietti *et al.* (2022). Our final spaces are made of respectively 1,954 verbs for `nsubj` and 787 verbs for `dobj`.

[5] `https://huggingface.co/distilbert-base-uncased`.
[6] `https://huggingface.co/gpt2-xl`.
[7] `https://huggingface.co/EleutherAI/pythia-70m`.

3.3 Learning algorithm

The experimental procedure, illustrated as follows, is repeated for each given model. As a first experiment, we trained a machine learning model to learn a linear mapping between the embedding space generated by our model and the target rating-based space previously created. The learning algorithm we used is a Partial Least Squares Regression with $k = 10$, in a ten-fold cross-validation setting using the scikit-learn Python library (Pedregosa *et al.* 2011).

After the creation of the target semantics space and the reconstruction of the raw sentences from the EWT, these settings have been deployed in a three-step procedure consisting of: i) feeding the sentences to the model and extracting the target verb embeddings; ii) running the linear mapping between the yielded model's representations and the human-ratings target space; iii) comparing the output yielded by this mapping with the original vectors derived from human judgements, using the Spearman rho correlation coefficient.

To assess the non-randomness of the process and the results, we generated random matrices shaped as the embedding spaces yielded by the model under consideration and replicated steps ii) and iii). For example, the performance of the embedding space generated by Pythia, having shape $1,954 \times 512$ corresponding to the number of verbs by the dimensions of each embedding, is compared with the performance of a matrix of the same shape $(1,954 \times 512)$ populated with random values sorted from the interval [0,1], which acts as a simulation of random embeddings. The comparison through correlation is done both at the row level pointing to verb vectors correlations, and at the column level, taking into account property vectors. Additionally, following the insight by Proietti *et al.* (2022), we applied a Sparse Principal Component Analysis (Mairal *et al.* 2009) reduction to the target spaces in order to reduce some background noise.

3.4 Causative-inchoative alternation prediction

In the second experiment, we aim to test whether the embeddings of the language models capture the causative-inchoative alternation. There is a simple assumption underlying this experiment, namely that, in verbs participating in the causative-inchoative alternation, the first argument will bear more Proto-Agent properties in a transitive context, while having more Proto-Patient properties in an intransitive one. To assess this hypothesis, we selected from the semantic spaces built for the `nsubj` only the transitive verbs, which are those with the highest expected Proto-Agent properties for their first argument. The resulting $1,854 \times 14$ space has been used to train regression models with the same algorithm of the previous experiment. Finally, the regressors have been

tested to predict a set of novel unseen verbs admitting the causative-inchoative alternation. The test set is the collection of sentences gathered by Proietti *et al.* (2022) containing 50 verb types participating in the causative-inchoative alternation, distributed in 100 sentences alternating between the transitive use and the intransitive use of the same target verb. This set has been manually crafted collecting sentences from sources such as VerbNet (Schuler 2006), FrameNet (Baker *et al.* 1998) and the enTenTen corpus via SketchEngine (Kilgarriff *et al.* 2014), while the prior selection of target verbs follows the criteria posed in the classification by Levin (1993). An exemplar subset of the verbs used is shown below in Section 4.2. Like in the first experiment, the same procedure was repeated for each model.

4. EXPERIMENTS AND RESULTS

4.1 Linear mapping and correlations

As described above, the first experiment consisted in running a linear mapping between an embedding space yielded by the language models and its rating-based counterpart. The Spearman correlation coefficient ρ was used to evaluate the generated output against the original vectors derived by human judgments. The aim is to understand to what extent the representation produced by language models encode Proto-Roles information: The higher the correlation, the better a certain property is represented in the model's embeddings.

We divided the analysis of the average ρ obtained by computing correlations at the verb level and the property level for both the `nsubj` and the `dobj` argument, from an inspection of single properties average correlations, again for both arguments.

Concerning the evaluation of average correlation values, the performance is slightly variable across models, but we observed a general tendency to replicate the results in Proietti *et al.* (2022) in that the average ρ is higher for the `nsubj` space than for the `dobj`. Overall, BERT seems to stay on top as the best-performing model in terms of average correlations. This is noticeable mostly in the comparison along the dimensions of the `nsubj` argument and the average correlations by verb level, as shown in Table 2. However, it is closely followed by its smaller counterpart, the DistilBERT model. Indeed, not only does the latter equal the former in the average correlation by property for the `dobj` but also slightly overtakes it in a specific spot, namely the average correlation by property for the `nsubj` argument, as can be seen in Table 2. Generally speaking, differently from BERT, all the other models have shown a negligible gap in the obtained average correlations verb-wise and property-wise for both the

nsubj and the dobj arguments to the point that there is no difference at all in the comparison between the two dimensions for DistilBERT, GPT2 and Pythia for the dobj.

The analysis got more complex when focusing on the examination of correlations by property and the evaluation of the ability of the employed methodology to model this phenomenon at such a fine-grained level. Taking once again the average of each model, each sub-space (nsubj/dobj) and each property, we found that there are differences in the goodness of the modelling at this level of analysis. That is to say that on the one hand, certain TLMs are better than others on average, on the other hand, they show a narrower gap at the level of single properties. For what concerns the average modelling of properties for the dobj argument, we registered a greater gap between the encoders and the decoders, with the former being the best-performing ones. Moreover, the two decoder models seem to exhibit a greater distance between the values obtained for the nsubj and those of the dobj arguments at both the verb and property levels. A more in-depth analysis of such resulting data will be provided in Section 5 where a plausible interpretation will be attempted.

The baseline for each model is recurrently equal to $\rho \sim 0$ at each level of analysis and for each sub-space, thus confirming the non-randomness of the reported values.

	nsubj		dobj	
	by verb	by property	by verb	by property
BERT	**0.51**	0.43	**0.40**	**0.38**
DistilBERT	0.46	**0.44**	0.38	**0.38**
GPT2-XL	0.42	0.41	0.26	0.26
Pythia70m	0.41	0.40	0.21	0.21

TABLE 2: COMPLETE VIEW OF THE AVERAGE SPEARMAN ρ FOR EACH MODEL ALONG EVERY DIMENSIONS.

As we noted before, the analysis is harder to interpret at the level of single properties correlation. Nonetheless, it is possible to make some partial generalizations about the behaviour of all models regarding the ability to simulate certain properties with our methods.

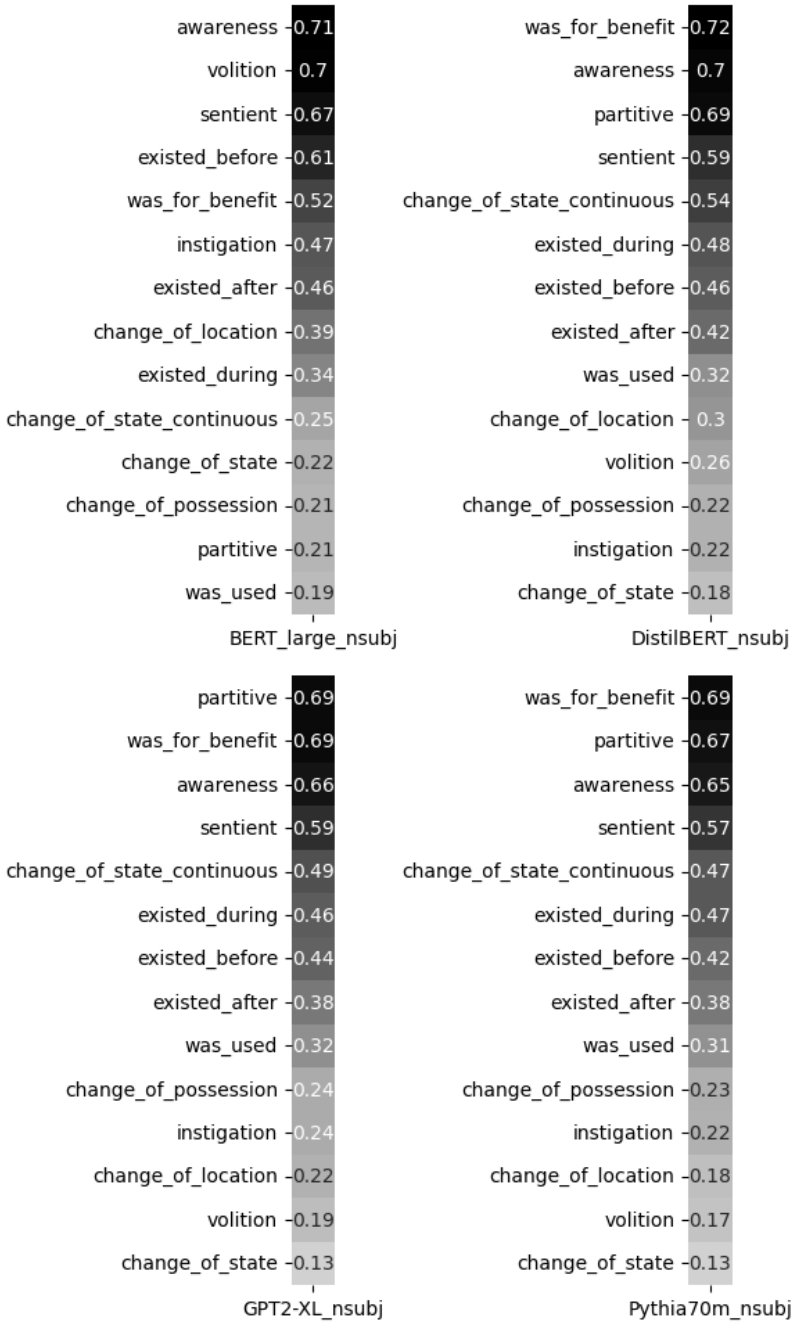

FIGURE 1: AVERAGE VALUES OF SINGLE PROPERTIES OBTAINED BY EACH MODEL FOR THE NSUBJ SORTED IN DESCENDING ORDER.

Interestingly but strangely, these generalizations are more stringent for the three models tested in the present paper, setting them a little aside in that respect from BERT's results. In fact, as can be seen in Figure 1 DistilBERT, GPT2-XL and Pythia seem to follow strikingly similar patterns of correlation quality with minor differences among the three models.

For example, regarding the `nsubj` argument, the `change of state` property results in being the hardest one to cope with for all these three models, scoring recurrently the lowest values. While `partitive` is the worst-modelled property for BERT, it reaches some of the highest values among the other models. A property which seems to be well-modelled for all the four models in the `nsubj` space is `awareness` along with `was for benefit`, in particular for DistilBERT, GPT2 and Pythia, and `sentient` for BERT. The central positions of the correlation scale for the `nsubj` are occupied by the "existential properties", `existed before`, `existed after`, `existed during` for all models except BERT. Another difference concerns the `volition` property, which is among the highest in BERT but gave poor results for all other models. A similar behaviour is that of `instigation`, which has a moderately high correlation value in BERT but seems to be difficult to model with any other model.

Observing the lower portion of Figure 1, it is possible to note that GPT2-XL and Pythia follow an almost identical indeed, with minor differences, mostly regarding the precise values obtained by each property. A tendency which is likely to be related to the shared architectural features among these models, which in fact differ significantly in size but are rooted in a common decoder architecture. However, this similarity is not replicated between BERT and DistilBERT, which on the other side share a similar encoder architecture.

Concerning the `dobj` argument, we got generally lower scores, but also even more mixed and less consistent values across all models. In fact, while for the `nsubj` we observed some parallelisms among the models in the way we were able to model single properties, as shown in Figure 2, for the direct object argument there is much more variability. Nonetheless, we can again individuate two consistent extremes across all models, namely the one of the best-modelled property, that is `change of possession`, and the one of the worst-modelled, which is, as it was for the `nsubj`, `change of state`.

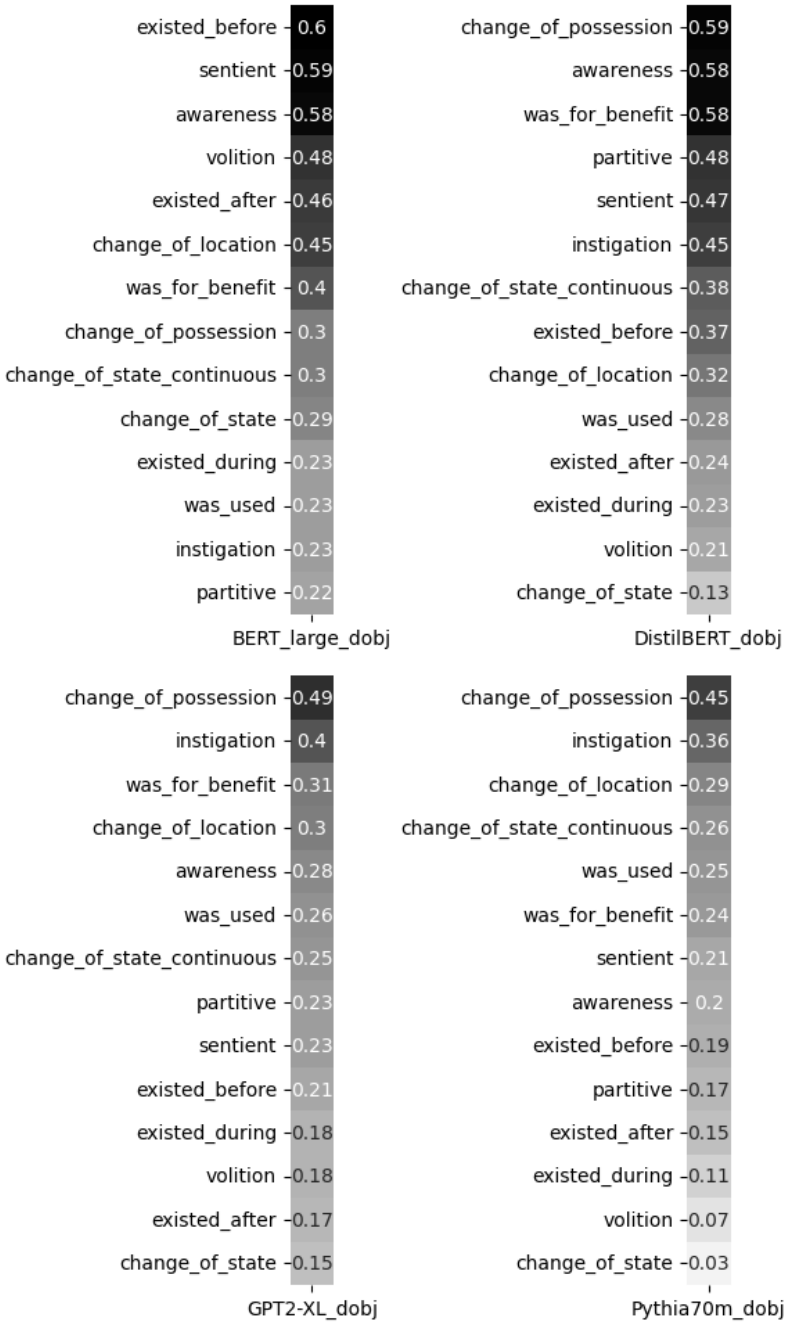

FIGURE 2: AVERAGE VALUES OF SINGLE PROPERTIES OBTAINED BY EACH MODEL FOR THE DOBJ SORTED IN DESCENDING ORDER.

4.2 Inchoative-causative alternation prediction

The second test was focused on modelling the causative-inchoative alternation through a prediction experiment, following the specifications described above. Having previously trained a regressor for each model, we tried to predict the properties of a set of 100 novel verbs participating in the causative-inchoative alternation. The set of verbs used comprises some of those listed by Levin (1993: Section 1.1.2). A sample is given in Table 3 where 25 of those verbs are shown grouped according to their corresponding VerbNet class.

Verb Class in VerbNet	Examples
break-45.1	*break, dissolve, shatter split, fracture*
bend-45.2	*bend, flex, roll, fold, drift*
cooking-45.3	*bake, dry, melt, cook, boil*
dress-41.1.1	*bathe, exercise, slide move, turn*
other_cos-45.4	*alter, blur, crumble degenerate, strengthen*

TABLE 3: EXAMPLES OF THE VERBS USED AND THEIR VERBNET CLASS. FIVE EXAMPLES PER CLASS.

For such verbs, we expected to obtain higher prediction values for properties characterizing the PROTO-AGENT role in those instances which are used in a transitive context. For example, a verb like *to bake* appears in two sentences representing a transitive and intransitive frame. The actual sentence corresponding to the transitive framework is *Jennifer baked the potatoes*, as opposed to an intransitive one being *The potatoes baked*. In other words, in the prediction phase, we expect to see higher values scored for PROTO-AGENT properties for the transitive version of the same verb.

To assess the validity of such a hypothesis and get quantitative insights from the exploration of the results, we pushed further our analysis, isolating from the whole set of predictions the two sub-spaces of TRANSITIVE and INTRANSITIVE use. For each of these target sub-spaces, we aggregated the predicted score for each property to confront the average values obtained between the two different uses of the same verbs. We expected to see higher average values for properties pertaining to the PROTO-AGENT in the transitive subspace and, vice-versa, higher values for the properties tied to the PROTO-PATIENT thematic role in the intransitive subspace.

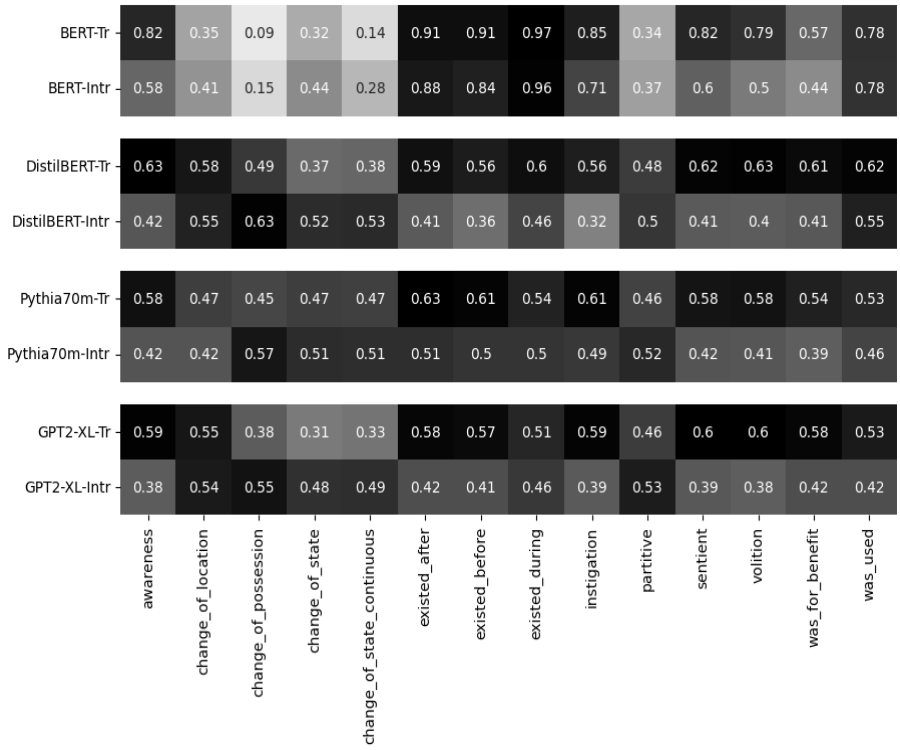

FIGURE 3: AVERAGE PREDICTIONS COMPARISON BETWEEN THE TRANSITIVE AND THE INTRANSITIVE SUB-SPACES FOR EACH MODEL.

Once again, we got mixed results, with a certain extent of variability across models and features. However, we were able to extract some consistent patterns from the analysis of single property average values in the comparison between the two spaces. As expected, we observed that the prediction yielded for each property has a tendency to follow our research hypothesis, assigning higher values to PROTO-AGENT properties when predicting transitive verbs. This can be seen in Figure 3, where properties like `awareness`, `was for benefit`, `volition`, `instigation` and `sentient` constantly obtained higher average scores for the transitive subspace across different models. On the other side, properties pertaining to the PROTO-PATIENT are generally scored higher in the intransitive space of each model, but this happens in a less strong manner. That is the case of entailments like `change of state`, `change of state continuous`, `partitive`. A property we failed to predict, with the embeddings of all models, is `was used`, for which our models assign constantly higher values to the transitive contexts, even if that one is theoretically a PROTO-AGENT property. Moreover, while BERT got equal average values

between transitive and intransitive, thus failing to obtain a clear-cut orientation toward one or the other subspace, the prediction obtained with the three other models assigned higher average values to transitive frames for that property. These findings may hide an intrinsic difficulty in modelling such property. Contrary to BERT, for which `change of location` was predicted on average with higher values for intransitive frames, thus considered a more PROTO-PATIENT oriented entailment in spite being theoretically a PROTO-AGENT property, all the other models treated here go in the opposite direction. In fact, we got higher average values for the transitive over their intransitive counterpart sub-space concerning that property both for DistilBERT and Pythia and GPT, as can be seen in Figure 3.

5. DISCUSSION

The results obtained through the experiments gave us the opportunity to make some interesting considerations about the ability of the tested models to encode proto-role information in their verb representations. First of all, it has been possible to recover this kind of information from each one of the employed models with a certain level of variability, which is different between the two arguments. The range of the obtained average correlation values spanned from $\rho = 0.41$ to $\rho = 0.51$ for the `nsubj` argument across all models, and from $\rho = 0.21$ to $\rho = 0.38$ for the `dobj` argument, taking into account the average verb correlations. Similarly, for the property-wise analysis, we got a range of average correlations going from $\rho = 0.40$ to $\rho = 0.44$, for the `nsubj` and from $\rho = 0.21$ to $\rho = 0.38$ for the `dobj`.

This variability is much smaller between the two GPT-like architectures tested, GPT2-XL and Pythia, and grows higher when we compare them with the two encoder-only models taken into account, BERT and DistilBERT. Such observations prompt us to make two primary considerations.

Firstly, the Pythia model, which has just 70m parameters, and the GPT2-XL, which is much bigger reaching 1.5B parameters, do perform differently, with a predictable advantage for the bigger model. However, such a difference in performance does not seem to mirror proportionally the great size leap between the two models. The GPT2-XL model is $\sim$21,5 times bigger than Pythia70M and the differences between the two in terms of correlations in the first experiment are in the order of the second decimal place, in some cases being of just a 0.01 of distance. Other than that, in the second experiment, as described in Section 4 and showed in Figure 1, correlation results for GPT2 and Pythia follow almost the same pattern in modelling the single properties for the `nsubj` argument with punctual differences of a really negligible amount. This

seems to suggest that size does not play a crucial role in the skill of the model to encode proto-role information. To confirm that hypothesis, we got DistilBERT, another small model, as the second best-performing one with $\sim$ 60M parameters. Moreover, comparing DistilBERT's results with those obtained with BERT we can see that the performances of the former smaller model are directly comparable with those of the latter bigger one, which are indeed equal or slightly better in some cases (see Table 2). In fact, while for the `nsubj` BERT got a correlation value by row of $\rho = 0.51$ for BERT, which is higher than the value obtained for DistilBERT ($\rho = 0.46$), the result at the column level of $\rho = 0.43$ is slightly overtaken by DistilBERT, which gave an average correlation of $\rho = 0.46$. Differences between the two models are even less marked concerning the `dobj` argument, with BERT giving values of $\rho = 0.40$ by row and $\rho = 0.38$ by column, against DistilBERT, which yield a $\rho = 0.38$ both by row and by column, loosing a 0.02 in the first case and performing equally to BERT in the second.

The second thought-provoking point regards the evaluation of two different architectures and their relative performances. In this respect, we can confidently say that, with the methods and data we used, BERT-like architectures seem to be more capable of encoding proto-role information in verb embeddings and generating better representations of such semantic knowledge in their spaces, as opposed to GPT-like models. Not only are the results of BERT still the overall best among the tested models, across most of the considered dimensions and levels of analysis, but DistilBERT, which follows closely its largest cousin in performance, is the second-best performing in the present study against the two GPT-like architectures employed, Pythia and GPT2-XL, despite being a much smaller size than GPT2-XL.

This is likely due to the essence of the architectures themselves. Encoder-only architectures like BERT and DistilBERT are primarily thought of as tools to build embeddings of lexical items with bidirectional context awareness. On the other side, GPT2-XL and Pythia are decoder-only architectures, which are trained with the main purpose of generating text in an autoregressive fashion, due to which they have partial knowledge of the context, focusing only on the previous/left side of a target token in a given sentence This perspective may make more sense if we focus our analysis on the `dobj`. This is the argument for which we got the larger span of distance in correlations between the GPT-like models ($\rho = 0.21/\rho = 0.26$) on one side and DistilBERT on the other side ($\rho = 0.38$). If we consider that in English sentences the `dobj` is likely to be on the right of the verb, we can expect that, due to their autoregressive nature, the GPT-like models are less aware of the direct object when generating the embedding of the verb, which entirely depends on the preceding elements,

including the subject. Therefore, we can expect that Proto-Role information about the direct object is less encoded in the verb embeddings yielded by generative models.

As described in Section 4, the second experiment was focused on the prediction of single properties for verbs participating in the causative-inchoative alternation and Figure 3 shows the results of such process. While it seems possible to predict such information with the embeddings of all four models to a certain extent, we wanted also to get insights about the differences across models in the quality of the prediction. Looking just at Figure 3 is possible to notice that in absolute terms, BERT got the highest prediction scores followed by DistilBERT, while the embeddings of the two decoder models seem to give very similar scores. However, to get a more relative evaluation across models we computed the differences between the average predicted properties for the transitive and the intransitive sub-spaces, thus having a sort of normalization across models. In doing so, the assumption is that whatever the absolute values yielded by the single model, a greater difference in the prediction of a given property should depict a greater ability to discriminate between the two averaged sub-spaces and eventually between the PROTO-AGENT and PROTO-PATIENT roles. That is shown in Figure 4, where a positive number means a prediction skewed toward the transitive subspace (i.e., transitive higher than intransitive), while a negative one means it is skewed toward the intransitive subspace (i.e., intransitive higher than transitive). In other words, a score, positive or negative, gave us a measure of how much a certain property has been valued for a certain space.

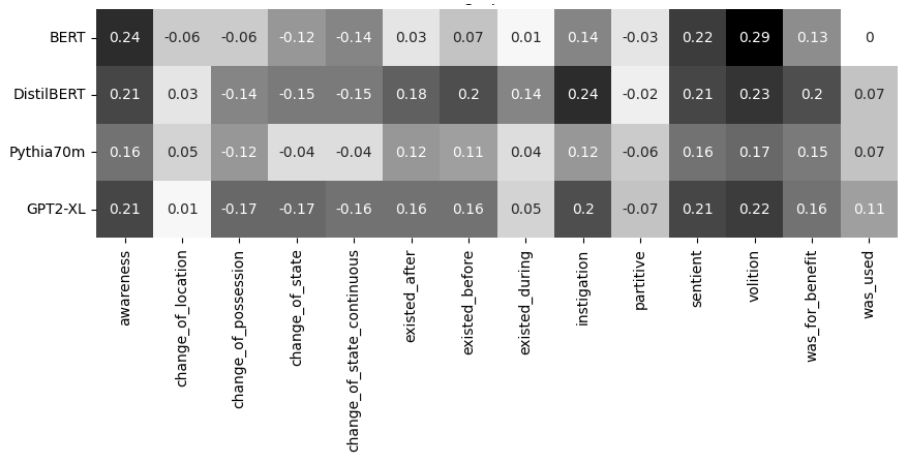

FIGURE 4: RELATIVE DIFFERENCES BETWEEN THE AVERAGE TRANSITIVE AND INTRANSITIVE PREDICTIONS FOR EACH MODEL.

For example, considering the property `awareness` we can see that it has been predicted with a higher average value of 0.24 in favor of the transitive subspace with BERT embeddings, 0.21 with those of DistilBERT and GPT2-XL and 0.16 with those of Pythia. Instead, a property like `change of possession`, has negative values for all models representing how much on average it has been predicted higher for the intransitive subspace. Looking at these values, prediction performances may be regarded differently. First of all, there is no notable difference among models, except for a few properties. For example, all three existential properties are very poorly discriminated with BERT embeddings, while they are more clearly skewed toward the transitive sub-spaces for the other models. Two properties which seemed to do in opposite directions between BERT and the other models are now more levelled with this visualization. In fact, a property like `change of location` which seemed skewed toward intransitive verbs for BERT, while toward transitive for the others, is in reality very similarly (bad-)discriminated among all models, the differences being in the order of the second decimal place. A similar account has to be done for `was used`, for which the difference is 0 for BERT, but around 0 for every other model. Another fact that may be worth pointing out is that `change of state` and `change of state continuous`, for which Pythia seem to struggle more than the other models in assigning higher values to the intransitive subspace for this entailment. Apart from the points briefly discussed above, Figure 4 shows a pretty homogeneous table of values with slight variability across the models and, all in all, the predictions with the different embeddings gave similar relative results. For example, while BERT is the best in some predictions following this metric, like in `awareness`, `sentient`, `volition` other models perform better in discriminating other property, like DistilBERT for `was for benefit` or GPT2-XL for `change of state`. This makes it difficult to individuate a clear best-performing model over the others, because that may change depending on the property considered.

6. CONCLUSIONS

All in all the results we obtained through our experiments are on track with those of Proietti *et al.* (2022) confirming that Transformer language models do encode some information about the semantic properties of the thematic proto-role held by the arguments of the verb, attesting this finding also for GPT-like architectures. Through the present work, we have been able to gain insights into the differences in performance between models built with different architectures and trained with different objective functions. In that respect, we noted that in the comparison between encoder-only and decoder-only architectures,

the former seem to have a better ability to represent information about proto-role's entailments in the verb embeddings. Additionally, confronting models with similar architectures but significantly different sizes, we found that for the sake of the knowledge targeted in this work, an increase in size does not mean a proportional gain in performance. Furthermore, by analysing the average prediction of properties of novel verbs participating in the so-called causative-inchoative alternation, we found that all models seem to be able to discriminate between PROTO-AGENT subjects, prototypical of the transitive version of those verbs, from PROTO-PATIENT subjects, realized on the contrary as the first argument of the intransitive framework of alternating verbs. Even though we noted some variability across models in such predictions, with the encoders yielding higher absolute values, we found that, by applying a sort of normalization to that output, all models seem to behave really similarly in relative terms.

Such findings notwithstanding, some limitations remain open and should be addressed in future works. On the methodology side, more experimentation is needed in trying to find a better learning algorithm to run the mapping, as we tested a single linear regression algorithm to infer new vectors and elicit correlation values. In that respect, other machine-learning methods or more complex, non-linear deep-learning ones may be attempted to accomplish such tasks and yield better mappings. On the model side, we did not test the newest generation of Transformer language models, the so-called *Large Language Models*, like GPT3 (Brown *et al.* 2020), Bloom (BigScience Workshop 2023), Llama (Touvron *et al.* 2023) and alike. These models are significantly larger than the ones we used here, surpassing the latter by several orders of magnitude and being clearly better in downstream tasks performances. The insights we gained and exposed in the present paper about the quality of the encoding of that proto-roles information at size scaling may change when taking into account these giant models. However, we see these limitations as more relevant for the technical side than the theoretical one, which is the main goal of the study. In fact, it appears that the verb embeddings yielded by language models do encode some information regarding the proto-role underlying the verb's arguments and do so in a way which seems to be changing more depending on the model's architecture than its size.

REFERENCES

Baker, C.F., C.J. Fillmore & J.B. Lowe (1998). The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*. Montreal, Quebec, Canada: Association for Computational Linguistics, 86–90. https://aclanthology.org/P98-1013.

Baroni, M. (2022). On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. In S. Lappin & J.P. Bernardy (eds.) *Algebraic Structures in Natural Language*, chapter 1, 1–16. CRC Press. https://doi.org/10.1201/9781003205388.

Biderman, S., H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M.A. Khan, S. Purohit, U.S. Prashanth, E. Raff, A. Skowron, L. Sutawika & O. van der Wal (2023). Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the International Conference on Machine Learning*. Honolulu, Hawaii, 2397–2430. https://proceedings.mlr.press/v202/biderman23a.html.

BigScience Workshop (2023). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.

Binder, J.R., L.L. Conant, C.J. Humphries, L. Fernandino, S.B. Simons, M. Aguilar & R.H. Desai (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33(3-4). 130–174. https://doi.org/10.1080/02643294.2016.1147426.

Bonial, C., W. Corvey, M. Palmer, V.V. Petukhova & H. Bunt (2011). A Hierarchical Unification of LIRICS and VerbNet semantic roles. In *Proceedings of the 5th IEEE International Conference on Semantic Computing*. Palo Alto, CA, USA, 483–489. https://doi.org/10.1109/ICSC.2011.57.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever & D. Amodei (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33. 1877–1901.

Burzio, L. (1986). *Italian syntax: A government-binding approach*. Dordrecht, Holland: Springer Science & Business Media.

Chersoni, E., E. Santus, C.R. Huang & A. Lenci (2021). Decoding Word Embeddings with Brain-Based Semantic Features. *Computational Linguistics*. https://doi.org/10.1162/coli_a_00412.

Cho, K., B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk & Y. Bengio (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In A. Moschitti, B. Pang & W. Daelemans (eds.) *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 1724–1734. https://aclanthology.org/D14-1179.

Conneau, A., G. Kruszewski, G. Lample, L. Barrault & M. Baroni (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2126–2136. https://aclanthology.org/P18-1198.

de Marneffe, M.C., C.D. Manning, J. Nivre & D. Zeman (2021). Universal Dependencies. *Computational Linguistics*, 47(2). 255–308. https://doi.org/10.1162/coli_a_00402.

de Varda, A. & M. Marelli (2023). Scaling in cognitive modelling: a multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Toronto, Canada: Association for Computational Linguistics, 139–149. https://aclanthology.org/2023.acl-short.14.

Devlin, J., M.W. Chang, K. Lee & K. Toutanova (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Minnesota, 4171–4186. https://www.aclweb.org/anthology/N19-1423.

Dowty, D.R. (1989). On the Semantic Content of the Notion of 'Thematic Role'. In G. Chierchia, B.H. Partee & R. Turner (eds.) *Properties, Types and Meaning: Volume II: Semantic Issues*, 69–129. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-2723-0_3.

Dowty, D.R. (1991). Thematic Proto-roles and Argument Selection. *Language*, 67. 547–619. https://www.jstor.org/stable/pdf/415037.pdf.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 55–65. https://aclanthology.org/D19-1006.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8. 34–48. https://aclanthology.org/2020.tacl-1.3.

Ferretti, T.R., K. McRae & A. Hatherell (2001). Integrating Verbs, Situation Schemas, and Thematic Role Concepts. *Journal of Memory and Language*, 44(4). 516–547. http://www.sciencedirect.com/science/article/pii/S0749596X00927281.

Făgărășan, L., E.M. Vecchi & S. Clark (2015). From Distributional Semantics to Feature Norms: Grounding Semantic Models in Human Perceptual Data. In *Proceedings of the 11th International Conference on Computational Semantics*. London, UK, 52–57. https://aclanthology.org/W15-0107.

Hare, M., M. Jones, C. Thomson, S. Kelly & K. McRae (2009). Activating event knowledge. *Cognition*, 111(2). 151–167. http://www.sciencedirect.com/science/article/pii/S0010027709000389.

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar*. Cambridge, MA: The MIT Press.

Kako, E. (2006a). The semantics of syntactic frames. *Language and Cognitive Processes*, 21(5). 562–575. https://doi.org/10.1080/01690960500101967.

Kako, E. (2006b). Thematic role properties of subjects and objects. *Cognition*, 101(1). 1–42. https://doi.org/10.1016/j.cognition.2005.08.002.

Kaplan, J., S. McCandlish, T. Henighan, T.B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu & D. Amodei (2020). Scaling laws for neural language models.

Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý & V. Suchomel (2014). The Sketch Engine: ten years on. *Lexicography*. 7–36. https://doi.org/10.1007/s40607-014-0009-9.

Lebani, G.E. & A. Lenci (2018). A Distributional Model of Verb-Specific Semantic Roles Inferences. In T. Poibeau & A. Villavicencio (eds.) *Language, Cognition, and Computational Models*, chapter 6, 118–158. Cambridge University Press. https://doi.org/10.1017/9781316676974.006.

Lebani, G.E. & A. Lenci (2021). Investigating dowty's proto-roles with embeddings. *Lingue e Linguaggio*, 20. 165–197. https://doi.org/http://dx.doi.org/10.1418/102812.

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4(1). 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254.

Lenci, A. (2023). Understanding natural language understanding systems. a critical analysis. *Sistemi Intelligenti*, 35(2). 277–302.

Lenci, A. & M. Sahlgren (2023). *Distributional Semantics*. Studies in Natural Language Processing, Cambridge University Press. https://doi.org/10.1017/9780511783692.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago, IL: University of Chicago press.

Levin, B. (2015). Semantics and pragmatics of argument alternations. *Annual Review of Linguistics*, 1(1). 63–83. https://doi.org/10.1146/annurev-linguist-030514-125141.

Levin, B. & M. Rappaport Hovav (2005). *Argument Realization*. Research Surveys in Linguistics, Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511610479.

Levin, B., M. Rappaport Hovav & S.J. Keyser (1995). *Unaccusativity: At the syntax-lexical semantics interface*. Cambridge, MA: The MIT press.

Liu, Q., M.J. Kusner & P. Blunsom (2020). A survey on contextual embeddings. *ArXiv*, abs/2003.07278. https://api.semanticscholar.org/CorpusID:212725577.

Mairal, J., F. Bach, J. Ponce & G. Sapiro (2009). Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09, New York, NY, USA: Association for Computing Machinery, 689–696. https://doi.org/10.1145/1553374.1553463.

McRae, K., T.R. Ferretti & L. Amyote (1997). Thematic Roles as Verb-specific Concepts. *Language and Cognitive Processes*, 12(2-3). 137–176. https://doi.org/10.1080/016909697386835.

McRae, K., M. Hare, J.L. Elman & T.R. Ferretti (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, 33(7). 1174–1184. https://link.springer.com/article/10.3758/BF03193221.

Mikolov, T., K. Chen, G. Corrado & J. Dean (2013a). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the 1st International Conference on Learning Representations, Workshop Track Proceedings*. Scottsdale, Arizona.

Mikolov, T., Q.V. Le & I. Sutskever (2013b). Exploiting Similarities among Languages for Machine Translation. *arXiv preprint arXiv:1309.4168*. https://arxiv.org/abs/1309.4168.

Oh, B.D. & W. Schuler (2022). Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 9324–9334. https://aclanthology.org/2022.emnlp-main.632.

Oh, B.D. & W. Schuler (2023). Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, 11. 336–350. https://doi.org/10.1162/tacl_a_00548.

Palmer, M., D. Gildea & P. Kingsbury (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1). 71–106. https://www.aclweb.org/anthology/J05-1004.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & E. Duchesnay (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12. 2825–2830. https://doi.org/10.5555/1953048.2078195.

Perlmutter, D.M. (1978). Impersonal passives and the unaccusative hypothesis. In *Annual Meeting of the Berkeley Linguistics Society*, volume 4. 157–190.

Proietti, M., G. Lebani & A. Lenci (2022). Does BERT Recognize an Agent? Modeling Dowty's Proto-Roles with Contextual Embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 4101–4112. https://aclanthology.org/2022.coling-1.360.

Pustejovsky, J. & O. Batiukova (2019). *The Lexicon*. Cambridge Textbooks in Linguistics, Cambridge, UK: Cambridge University Press.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei & I. Sutskever (2019). Language Models are Unsupervised Multitask Learners. *OpenAI blog*. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Reisinger, D., R. Rudinger, F. Ferraro, C. Harman, K. Rawlins & B. Van Durme (2015). Semantic Proto-Roles. *Transactions of the Association for Computational Linguistics*, 3. 475–488. https://www.aclweb.org/anthology/Q15-1034.

Rogers, A., O. Kovaleva & A. Rumshisky (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8. 842–866. https://aclanthology.org/2020.tacl-1.54.

Rosch, E. & C.B. Mervis (1975). Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*, 7. 573–605. https://doi.org/10.1016/0010-0285(75)90024-9.

Rudinger, R., A. Teichert, R. Culkin, S. Zhang & B. Van Durme (2018). Neural-Davidsonian Semantic Proto-role Labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, 944–955. https://www.aclweb.org/anthology/D18-1114.

Sanh, V., L. Debut, J. Chaumond & T. Wolf (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108. https://api.semanticscholar.org/CorpusID:203626972.

Schuler, K.K. (2006). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania. http://verbs.colorado.edu/~kipper/Papers/dissertation.pdf.

Schwartz, D. & T. Mitchell (2019). Understanding language-elicited EEG data by predicting it from a fine-tuned language model. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 43–57. https://aclanthology.org/N19-1005.

Silveira, N., T. Dozat, M.C. de Marneffe, S. Bowman, M. Connor, J. Bauer & C.D. Manning (2014). A Gold Standard Dependency Corpus for English. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*. Reykjavik, Iceland, 2897–2904. http://www.lrec-conf.org/proceedings/lrec2014/summaries/1089.html.

Stengel-Eskin, E., A.S. White, S. Zhang & B. Van Durme (2020). Universal Decompositional Semantic Parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8427–8439. https://aclanthology.org/2020.acl-main.746.

Stengel-Eskin, E., K. Murray, S. Zhang, A.S. White & B. Van Durme (2021). Joint Universal Syntactic and Semantic Parsing. *Transactions of the Association for Computational Linguistics*, 9. 756–773. https://doi.org/10.1162/tacl_a_00396.

Sutskever, I., O. Vinyals & Q.V. Le (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence & K. Weinberger (eds.) *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave & G. Lample (2023). Llama: Open and efficient foundation language models.

Utsumi, A. (2020). Exploring What Is Encoded in Distributional Word Vectors: A Neurobiologically Motivated Analysis. *Cognitive Science*, 44(6). e12844. https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12844.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser & I. Polosukhin (2017). Attention is all you need. In I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (eds.) *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vulić, I., E.M. Ponti, R. Litschko, G. Glavaš & A. Korhonen (2020). Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 7222–7240. https://aclanthology.org/2020.emnlp-main.586.

White, A.S., D. Reisinger, K. Sakaguchi, T. Vieira, S. Zhang, R. Rudinger, K. Rawlins & B. Van Durme (2016). Universal Decompositional Semantics on Universal Dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, 1713–1723. https://www.aclweb.org/anthology/D16-1177.

White, A.S., E. Stengel-Eskin, S. Vashishtha, V.S. Govindarajan, D.A. Reisinger, T. Vieira, K. Sakaguchi, S. Zhang, F. Ferraro, R. Rudinger, K. Rawlins & B.V. Durme (2020). The Universal Decompositional Semantics Dataset and Decomp Toolkit. In *Proceedings of the 12th Conference on Language Resources and Evaluation*. Marseille, France, 5698–5707. https://www.aclweb.org/anthology/2020.lrec-1.699.pdf.

Mattia Proietti
University of Pisa – DiLLeS PhD Programme
via Santa Maria 36, Pisa
Italy
e-mail: mattia.proietti@phd.unipi.it
https://orcid.org/0009-0002-0447-680X

Gianluca E. Lebani
Ca' Foscari University of Venice
Department of Linguistics and Comparative Cultural Studies
Ca' Bembo, Fondamenta Tofetti, Dorsoduro 1075, Venice
Italy
e-mail: gianluca.lebani@unive.it
https://orcid.org/0000-0002-3588-1077

Alessandro Lenci
University of Pisa – Department of Philology, Literature, and Linguistics
via Santa Maria 36, Pisa
Italy
e-mail: alessandro.lenci@unipi.it
https://orcid.org/0000-0001-5790-4308