**Gianluca E. Lebani, Alessandro Lenci**

# Investigating Dowty's proto-roles with embeddings

**(doi: 10.1418/102812)**

Lingue e linguaggio (ISSN 1720-9331)
Fascicolo 2, luglio-dicembre 2021

**Ente di afferenza:**
*Università degli studi di Pisa (Unipi)*

# INVESTIGATING DOWTY'S PROTO-ROLES WITH EMBEDDINGS

GIANLUCA E. LEBANI ALESSANDRO LENCI

ABSTRACT: Distributional semantics represents words as multidimensional vectors recording their statistical distribution in context. Notwithstanding the wide use of this approach in fields as distant as Natural Language Processing, psycho-linguistic modeling and semantic analysis, relatively little work focused on the characterization of the semantic information encoded in these semantic vectors, especially for verbs. Here we investigate whether and to what extent distributional vectors are able to encode the semantic content of Dowty's semantic proto-roles, which can be characterized as the set of entailment relations that an argument receives by virtue of its role in the event described by a predicate (Dowty 1989, 1991). We created several linear mappings between various kinds of static embeddings and a semantic space built on the basis of the proto-roles annotations collected by White *et al.* (2016). Our results show that, to a certain extent, proto-roles information is available in distributional models, and that a linear mapping can be used to infer the semantic characteristics of the arguments of novel verbs, thus testing the possibility of developing large-scale models able to extract the semantic properties for a wide inventory of verbs. Finally, we report a qualitative analysis in which we discuss which entailment relations our technique associates with a few semantic verb classes whose semantic roles are notoriously difficult to describe.

KEYWORDS: Thematic proto-roles, semantic roles, distributional semantic models, word embeddings.

## 1. INTRODUCTION[1]

Distributional semantics, probably the most exciting technique developed in the last 20 years for the computational representation of word meaning, represents lexical items as multidimensional vectors (i.e. *word embeddings*) learned from the analysis of the statistical distribution of words in context (Lenci 2018; Boleda 2020). Distributional semantics is a usage-based approach based on the so-called distributional hypothesis (Harris 1954), according to which words

[1] This work is the result of a joint work by both authors. For the specific purposes of Italian Academy, Gianluca E. Lebani is responsible for Sections 3-7, and Alessandro Lenci for Sections 1-2. The authors would like to thank Aaron Steven White and an anonymous reviewer for commenting on an earlier version of this paper.

occurring in similar contexts tend to share some semantic traits. Under this view, the fact that words like *dog* and *cat* are semantically related can be inferred from the fact that they appear in similar contexts like *he was playing with his _*, *petting a _ is relaxing* and *pour some water in the _'s bowl*.

In the years, many different techniques have been proposed to build word embeddings. Such techniques, often referred to as Distributional Semantic Models (DSMs), can be characterized along many different dimensions (for an updated overview, see Lenci 2018; Lenci *et al.* 2021). In this context, probably the most important distinction is between the first-generation *count models* and the younger *predict models*, based on neural networks. Count models are called in this way because the vector for a given lexical element is built on the basis of its co-occurrence frequency, that is how often it co-occurs with a given set of contextual features. DSMs of this type include both classical works such as Hyperspace Analogue of Language (HAL: Lund & Burgess 1996) and Latent Semantic Analysis (LSA: Landauer & Dumais 1997), as well as more modern takes on the matter like GloVe (Pennington *et al.* 2014).

Predict DSMs, on the other way around, take a radically different approach, in which word vectors are obtained by training a artificial neural network to optimally predict the context of a set of target words. The most famous DSMs of this family are those implemented in the word2vec library (Mikolov *et al.* 2013a,c) and in its extension fastText (Bojanowski *et al.* 2017). Recently, a new strain of predict models emerged, that builds different vectors for each word token in context. The most famous architectures of this family include BERT (Devlin *et al.* 2019) along with all of its offspring, and GPT (Radford *et al.* 2019; Brown *et al.* 2020), and are often referred to as *contextualized DSMs*, in contrast with more traditional approaches that are sometimes labelled as *static embeddings*.

Word embeddings are widely used in many fields, from Text Mining to Natural Language Processing, from psycholinguistic modeling to theoretical linguistics and semantic analysis. In the development of NLP and AI application, indeed, it is very common to feed word embeddings to classifiers of any sort in order to provide semantic information that has proven to be useful for tasks as different as sentiment analysis (e.g., Yu *et al.* 2017; Kumar *et al.* 2021), hate-speech detection (e.g., Jain *et al.* 2021), word sense disambiguation (e.g., Loureiro *et al.* 2021) and machine translation (e.g., Qi *et al.* 2018). In linguistics, word vectors have been used to study topics as different as semantic change over time, polysemy, composition, morphology and various phenomena at the syntax–semantics interface (for a review, see Boleda 2020). From a cognitive point of view, DSMs have been shown to be apt to explain phenomena as diverse as semantic priming (e.g., Mandera *et al.* 2017), the strength of word associations (e.g., Jones *et al.* 2018; Utsumi 2015), the content of featural knowledge (e.g., Chersoni *et al.* 2021; Utsumi 2020; Făgărășan *et al.* 2015; Johns & Jones 2012; Riordan & Jones 2011) and the brain activity associated with semantic processing (e.g. Anderson *et al.* 2017; Huth *et al.* 2016; Mitchell *et al.* 2008).

Notwithstanding their wide use, relatively few works focused on characterizing the kind of semantic information that word embeddings encode. Probably the most widely studied kinds of information are the emotional information (e.g., Lenci *et al.* 2018; Passaro *et al.* 2017; Recchia & Louwerse 2015), attitudes and bias (e.g., Caliskan *et al.* 2017) and the taxonomic, perceptual and commonsense knowledge that is encoded in verbal feature norms like those collected by McRae *et al.* (2005a) (e.g., Rubinstein *et al.* 2015; Făgărășan *et al.* 2015; Johns & Jones 2012; Riordan & Jones 2011), those by Vinson & Vigliocco (2008) (e.g., Johns & Jones 2012), the CLBS norms by Devereux *et al.* (2014) (Sommerauer & Fokkens 2018), or by conducting an ad-hoc elicitation experiment (e.g., Grand *et al.* 2018). A wider spectrum of semantic information is taken into consideration by Chersoni *et al.* (2021) and Utsumi (2020), who focused on the sensory, motor, spatial, temporal, affective, social, and cognitive features rated in the Binder *et al.* (2016) norms.

The present work inscribes itself in this line of research, by investigating whether and to what extent static embeddings of English verbs built by different DSMs are able to encode the entailment relations that, according to Dowty (1989, 1991)'s approach, form the semantic content of a *proto-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, there are only a handful of works that addresses the feasibility of a distributional approach to model this semantic information. One of this works is Lebani & Lenci (2018), in which the authors developed a DSM able to represent the thematic role properties activated by a subset of English verbs and tested their approach against elicited norming data collected by using the paradigm proposed by Lebani *et al.* (2015). Other relevant studies are Rudinger *et al.* (2018) and Stengel-Eskin *et al.* (2020, 2021), in which contextual information is used to parse text into a semantic representation implementing Dowty's theory of thematic proto-roles, the Universal Decompositional Semantics (White *et al.* 2016).

To test our reference DSMs, we use a linear mapping strategy akin to those employed by Chersoni *et al.* (2021), Utsumi (2020),[2] and Făgărășan

[2] For the sake of completeness, it should be reported that Utsumi (2020) employed two different mapping strategies: a linear mapping and a multi-layer perceptron. Given the exploratory nature of this paper, we chose to adopt the strategy that needed the least number of hyper-

*et al.* (2015), that were in turn inspired by the linear mapping strategy used by Mikolov *et al.* (2013b) for completely different purposes. Mikolov *et al.* (2013b), indeed, noticed that similar cluster of words from different languages tend to arrange in a geometrically similar ways. For instance, the relative position of English numerals (*one* to *five*) were similar to those of their Spanish counterparts (*uno* to *cinco*), and the same appear to be true for animal names like *cat*, *horse*, *cow*, *pig* and *dog*. Moving from this observation, these scholars tested whether a simple linear transformation were able to capture the relationship between vectors.

Chersoni *et al.* (2021), Utsumi (2020), and Făgărăşan *et al.* (2015), on the other hand, focused on the feasibility of creating a linear mapping between different distributional spaces and a semantic space built from an already available dataset of speaker-elicited ratings or descriptions. Făgărăşan *et al.* (2015)'s goal was to test whether such a strategy could be used to generate, for novel unseen words, feature-based representations. These authors created several mappings from either the context-predicting word2vec pre-trained embeddings Mikolov *et al.* (2013a,c) or the embeddings created by several versions of the same count-based space, to a semantic space created from the McRae *et al.* (2005b) norms and tested, for each vector mapped in the semantic space, whether its neighbors corresponded to the neighbors of the featural vector. Utsumi (2020), on the other way around, was more interested in identifying the knowledge encoded in word vectors. To this purpose, he created several mappings between several static DSMs and a semantic space created from Binder *et al.* (2016)'s conceptual representations and analyzed the performance of the inferred space among various types of information. Chersoni *et al.* (2021) applied a similar method to a larger array of DSMs, including contextualized ones, and used the featural space mapped from embeddings to explain the performance of DSMs in classification tasks targeting several semantic dimensions of nouns and verbs. In his spirit, our work is inspired by Chersoni *et al.* (2021) and Utsumi (2020). In a similar fashion, indeed, we built several mappings between several static DSMs to a reference semantic space, in our case the proto-roles annotations collected by White *et al.* (2016), and estimated the similarity between the dimensions of the inferred space and those of the reference semantic space in order to understand if and to what extent proto-roles information is available in these models.

This paper is organized as follows. In Section 2, we review the notion of semantic proto-role, with a particular focus on how it has been received in the NLP literature. In sections 3 and 4, we describe our method and test it against

parameters to set, that is the linear one, leaving for future research the comparison between different mapping strategies.

the gold ratings by White *et al.* (2016). The remainder of the paper is devoted to a qualitative analysis of how our technique deals with two oppositions that are often accounted for in terms of semantic roles: that between unergative and unaccusative verbs and that between admire and amuse verbs.

## 2. SEMANTIC PROTO-ROLES

Both in the theoretical and in the applied linguistic tradition, the role played by an argument in the event or situation described by a predicate is traditionally described by resorting to the notion of *semantic role* (a.k.a. *case relation*, *thematic roles/relations*, *theta role* or *semantic relation*). In other words, a semantic role is a label associated to an argument by virtue of its role in the event or situation described by a predicate (Levin & Rappaport Hovav 2005: ch. 2). Canonical examples of semantic roles are AGENT, that can be defined as the *"animate and volitional initiator or doer of an action"*, and PATIENT/THEME, that can be defined as the *"entity undergoing the action and somehow affected by it"* (Pustejovsky & Batiukova 2019: 29).

This approach, which can be traced back to Pānini's *kārakas*, treats semantic roles as linguistic primitives describing a natural class of arguments and is problematic in many ways, as reviewed by Dowty (1991: 553-559). The biggest issues are probably the lack of consensus on how semantic roles should be defined, how fine-grained they should be and, how they should be identified (Levin & Rappaport Hovav 2005). For instance, in many situations or events it is difficult to decide which role is assigned to a given argument, as it is the case for *John* in the following sentence:

(1) *John ran into the house*

Is *John* a PATIENT, because he initiates the movement, a THEME, because he moves, or, as suggested by Jackendoff (1972), both? The latter solution, in turn, would violate the uniqueness assumption underlying most theories of thematic roles and would require a re-thinking of how semantic roles should be identified and described in these approaches.

In recent years, following the seminal work of Dowty (1989, 1991), a novel approach emerged, according to which thematic roles can be seen as a bundle of more primitive entities, thus mirroring the prototype structure of other types of concepts. In this view, a role is a set of properties or entailments imposed by the predicate over its arguments. Some of these entailments are *verb-specific*, in that they follow from the meaning of the verb, while others are said to be *linguistic* in that they more abstract properties that are licensed by many verbs.

Being interested in these shared entailments, Dowty (1991) identified the two clusters of linguistic properties in (2) and (3) that he labelled as the PROTO-AGENT and the PROTO-PATIENT:

(2) 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)

(3) 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)

In Dowty's view, PROTO-AGENTS and PROTO-PATIENTS tend to be realized in active sentences as subjects and objects, respectively, and are organized like the prototypes described by Rosch & Mervis (1975). Such a prototype structure accounts for the fact that an argument is not bound to receive all the entailments of a given proto-role, and that the *agent-hood* or *patient-hood* of an argument is a function of the number of PROTO-AGENT and PROTO-PATIENT entailments received from the predicate. For instance, the subject and object of the verb *to build* possess all the properties from (2) and (3), making them "good" examples of an AGENT and of a PATIENT, respectively (Dowty 1991: 572). Conversely, subjects of psych predicates like *to fear* appear to be less "agentive" in that they lack the PROTO-AGENT volitionality and causer entailments (Dowty 1991: 573). This is consistent with the observation that in a traditional approach this argument usually receives a different semantic role labelled EXPERIENCER, that can be defined as the *entity psychologically or emotionally affected by the event* (Pustejovsky & Batiukova 2019: 29).

## 2.1 Collections of proto-role entailments

Dowty (1991)'s view received empirical support from psycho-linguistic evidence (McRae *et al.* 1997; Ferretti *et al.* 2001; McRae *et al.* 2005b; Kako 2006a,b; Hare *et al.* 2009) as well as from works belonging to the Computational Linguistics field (Reisinger *et al.* 2015; Lebani & Lenci 2018). In the

NLP community, Van Durme, Rawlins and colleagues even proposed a novel task, Semantic Proto-Role Labelling, aimed at the annotation of a sentence with *"scalar judgments of Dowty inspired properties"*, as opposed to the more conventional Semantic Labelling task based on the more conventional thematic roles (Reisinger *et al.* 2015; Teichert *et al.* 2017; White *et al.* 2017).

Relevant for our work is the corpus-based verification of Dowty's theory conducted by Reisinger *et al.* (2015) and by White *et al.* (2016). These scholars, indeed, not only showed that the proto-role hypothesis holds true when tested on large scale corpus-based data, but they also collected two datasets of proto-roles annotations that are publicly available as part of the Universal Decompositional Semantics Dataset (White *et al.* 2020).[3] Inspired by the work of Kako (2006b), Reisinger *et al.* (2015) developed a crowd sourcing annotation task in which each annotator was presented with a sentence from a subset of PropBank (Palmer *et al.* 2005) with an highlighted argument. For each item, the task was to judge, on a 5-points Likert scale, the plausibility of a property of the highlighted argument by answering to a question of the form *"How likely or unlikely is it that ARG is sentient?"*. The questions submitted to the annotators were derived by inspecting the definition of the role hierarchy by Bonial *et al.* (2011) and selecting all the properties that were most similar to the original questions proposed by Dowty. For each argument token, the annotators answered to 12 different semantic questions thus describing the following role properties: *instigated*; *volitional*; *awareness*; *sentient*; *moved*; *physical existed*; *existed before*; *existed during*; *existed after*; *changed possession*; *change of state*; *stationary*. The authors report to have collected judgments for over 9,000 arguments of near 5,000 verb tokens, spanning 1,610 PropBank verb sense IDs (i.e., *rolesets*). White *et al.* (2016) further revised Reisinger *et al.* (2015)'s protocol in many ways. Relevant for our purposes are the improvement of the inventory of annotated properties, here reported in Table 1 and the use of redundant annotations. These authors used their revised protocol to annotate the Universal Dependencies English Web Treebank (version 1.2: Silveira *et al.* 2014), a corpus that covers a wider range of genres then those covered by Propbank and whose syntactic structure is annotated according to the Universal Dependencies (de Marneffe *et al.* 2021: UD) guidelines. An inspection of the published dataset shows that this improved protocol has been used to collected 206,018 annotations (198,002 involving a NP argument) for 957 verbs occurring in 2,793 sentences; 4,607 verbal tokens have been annotated (4,600 of which had at least one NP argument), for a total of 7,144 verb-argument pairs (6,142 if we consider nominal arguments only).

[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: ROLES OF THE REVISED STRATEGY DESCRIBED BY WHITE *et al.* (2016).

## 3. METHODS

To understand if and to what extent proto-roles are encoded in verb embeddings, we evaluated how accurately vector spaces built with different DSMs can simulate (a selection of) the proto-role properties collected by White *et al.* (2016). In the experiment described in this section, we built a linear mapping between different distributional spaces and a entailment-based space. Following Chersoni *et al.* (2021) and Utsumi (2020), performance will be assessed by comparing the vectors in the entailment-based space against those predicted from the distributional spaces.

### *3.1 Distributional spaces*

We experimented with a restricted number of reliable static DSMs all trained on the same corpus, a concatenation of ukWaC (Baroni *et al.* 2009) with a 2018 dump of the English Wikipedia[4] parsed with CoreNLP Manning *et al.* (2014). The corpus size of this corpus is just shy of 4B tokens, and its vocabulary size is approximately 15.3M types. In this study, we decided to ignore contextualized

[4] `https://dumps.wikimedia.org`

embedding for two main reasons. First of all, while all our spaces are trained on the same corpus, models like BERT and (Devlin *et al.* 2019) and GPT-3 (Brown *et al.* 2020) are trained on corpora that include also document types different from those that compose the corpus used for our static DMS (e.g., the BERT training corpus includes books). Moreover, contextualized embeddings produce an output that should be manipulated to make it comparable with the representation created by static embeddings, for instance by averaging contextualized embeddings as proposed by Bommasani *et al.* (2020). As a consequence, we felt that adding these sources of variability could potentially mine the generalizability of our results, so we chose to leave the analysis of contextual models to a follow-up study.

The selection of static DSMs in our experiments, along with the setting of the relevant hyper-parameters, was based on the results of the evaluation by Lenci *et al.* (2021), from which the actual vector spaces were borrowed. We ended up with twelve 300-dimensional vector spaces, resulting from the application of four very popular learning methods with different context types.

- SVD.* models are built by counting the co-occurrences between the target words and the top 10,000 most frequent lexemes, weighting the raw counts with the smoothed PPMI described by Levy *et al.* (2015) with $\alpha = 0.75$, and applying Singular Value Decomposition (SVD) to reduce the dimensionality of the vectors to 300. Four different spaces were prepared by implementing four different kinds of context:
  - `SVD.2w`: two words in a sentence are counted as co-occurring if their linear distance is less than or equal to 2.
  - `SVD.10w`: two words in a sentence are counted as co-occurring if their linear distance is less than or equal to 10.
  - `SVD.synf`: in this syntactic-filtered space, two words in a sentence are counted as co-occurring if there is a dependency relation holding between them, but the syntactic information is lost in the vector space: two contextual words linked to the target word by different relations are treated as instances of the same context (Padó & Lapata 2007).
  - `SVD.synt`: in this syntactic-typed space, two words in a sentence are counted as co-occurring if there is a dependency relation holding between them, and such syntactic information is explicitly encoded in the vector space: two contextual words linked to the target word by different relations are treated as instances of different contexts (Baroni & Lenci 2010).

- Global Vectors (*GloVe.**) spaces are build by applying a weighted linear regression to the co-occurrence counts, with the training objective of learning word vectors whose dot product equals the logarithm of the words' probability of co-occurrence (Pennington *et al.* 2014). Two spaces were created by using a symmetrical 2 words (`GloVe.2w`) or a symmetrical 10 words (`GloVe.10w`) contextual span.
- Skip-Gram with Negative Sampling (*SGNS.**) spaces are built by using a two-layer neural network to predict, for each word, its surrounding context (Mikolov *et al.* 2013c). Four different models were created by implementing the same kinds of contexts described for the `SVD.*` models above (`SGNS.2w`, `SGNS.10w`, `SGNS.synf` and `SGNS.synt`). All the spaces were created by using the word2vec implementation available in the word2vecf library by Levy & Goldberg (2014).
- *fastText.** spaces are built by using the extension of SGNS described by Bojanowski *et al.* (2017), in which word vectors are learned for character n-grams rather than for entire words. Two spaces were built by using a symmetrical 2 words (`fastText.2w`) or a symmetrical 10 words (`fastText.10w`) contextual span.

### 3.2 Target entailment-based space

The target vector space was derived from the properties annotated by White *et al.* (2016). We preferred this dataset to that collected by Reisinger *et al.* (2015) for three main reasons. Firstly, White *et al.* (2016) adopted an improved classification of entailments, obtained by removing some redundant properties and by adding three properties targeting new types of arguments. Secondly, the texts annotated in this dataset come from a wider set of genres than those annotated by Reisinger *et al.* (2015). Finally, White *et al.* (2016)'s dataset is based on a UD-annotated treebank, whose inventory of core syntactic relations is more expressive than the one adopted in the Propbank. As a consequence, each entry in this dataset is marked with one of the following labels describing the syntactic relation holding between the verb and the argument: `nsubj`, for nominal subject, `nsubjpass`, for passive nominal subject, `dobj`, for direct object, `iobj`, for indirect object, and `nmod`, for nominal modifier. On the other hand, in the available grammatical functions in Reisinger *et al.* (2015)'s dataset are subject, object and "other".

In order to avoid data sparsity and categorical ambiguity issues, a series of filters were applied to the White *et al.* (2016)'s ratings before encoding them in a vector space. In a first step, we removed the few annotations collected with the Reisinger *et al.* (2015)'s procedure, along with the annotations of the HITs that were marked as "not applicable" and all verb-argument tokens involving a PP argument. This left us with 150,010 judgments for 957 verbs. We then went on to filter out modals and auxiliaries, the verbs that were rated by less than 100 times, and the verbs that occurred less than 1,000 times in the training corpus of our distributional spaces. Since our DSMs are not PoS-tagged, thus conflating into a single vector the distributional information of lemmas that belong to different grammatical classes, we filtered out all the lemmas whose occurrences in the corpus were tagged as a verb less than 75% of the times (e.g. `to address`, `to date`, `to email`). This left us with 72,038 ratings for 156 verbs. Finally, we removed all the ratings referred to a $\langle grammatical\_function, property\rangle$ pair annotated less than 150 times. The latter filter resulted in the exclusion of all the ratings for the `iobj` arguments, along with the annotations of the `change_of_state_continuous` property for the `nsubjpass` arguments. Wrapping up, the whole filtering process left us with 70,512 annotations for the `nsubj`, `nsubjpass` and `dobj` arguments of 156 verbs.

We then moved to aggregate the data in order to obtain a vector space structured as shown in Table 2: a matrix composed of 156 verb vectors whose dimensions are the 41 $\langle grammatical function, property\rangle$ pairs that have been annotated more than 150 times. The inventory of the matrix dimensions include all the entailments in Table 1 for the `nsubj`, `nsubjpass` and `dobj` arguments, with the exception of the `change of state continuous` property for the `nsubjpass` arguments. The values that populates this matrix are obtained by first averaging over the redundant annotations (i.e., annotations of the same token by different subjects), then averaging over all the annotations for each verb in each selected $\langle grammatical function, property\rangle$ pair and finally scaling this aggregated scores to the range $[0, 1]$.

In intuitive terms, the scores populating our matrix measure how plausible is for a given argument (e.g., the subject) of a given verb (e.g., `to kill`) to possess a given property (e.g., `awareness`). For instance, the first two values of the first vector of the space in Table 1 encode the fact that the subject of the verb `to kill` is very high in awareness and volition, while the latter property has a very low value for the direct object of the same verb.

A major criticism that can be leveled to this representation is that, by averaging over all the actual contextualized human judgments, it overlooks the fact that semantic roles are often assigned contextually. However, it should be stressed that the vectors in our entailment-based semantic space are basically abstract representations encompassing all the semantic roles that can be associated with all the different uses of out target verbs. In a sense, this is not

| | ⟨nsubj, awareness⟩ | ⟨nsubj, change of state⟩ | ⟨nsubj, volition⟩ | ⟨nsubjpass, awareness⟩ | ⟨nsubjpass, change of state⟩ | ⟨nsubjpass, volition⟩ | ⟨dobj, awareness⟩ | ⟨dobj, change of state⟩ | ⟨dobj, volition⟩ |
|---|---|---|---|---|---|---|---|---|---|
| `affect` | 0 | 0.625 | 0 | 0.875 | 1 | 0.125 | 0.688 | 0.75 | 0.187 |
| `amaze` | 0.25 | 0.25 | 0.25 | 1 | 0.75 | 1 | 1 | 0.708 | 0.792 |
| `bring` | 0.922 | 0.422 | 0.828 | 0.5 | 0.187 | 0.5 | 0.562 | 0.472 | 0.319 |
| `fill` | 0.875 | 0.25 | 0.875 | 0.625 | 0.687 | 0.5 | 0.5 | 0.875 | 0.562 |
| `give` | 0.899 | 0.352 | 0.887 | 0.5 | 0.125 | 0.5 | 0.062 | 0.312 | 0.081 |
| `ignore` | 1 | 0.875 | 1 | 0.583 | 0.583 | 0.083 | 0.75 | 0.5 | 0.125 |
| `include` | 0.458 | 0.51 | 0.433 | 0.25 | 0.25 | 0.25 | 0.451 | 0.461 | 0.446 |
| `kill` | 0.925 | 0.65 | 0.875 | 0.594 | 1 | 0 | 0.575 | 0.937 | 0.042 |
| `put` | 0.833 | 0.492 | 0.84 | 0.2 | 0.458 | 0 | 0.275 | 0.75 | 0.11 |
| `tell` | 0.99 | 0.357 | 0.959 | 0.949 | 0.536 | 0.574 | 0.968 | 0.561 | 0.714 |
| `warn` | 1 | 0.25 | 0.958 | 1 | 0.375 | 0.5 | 0.875 | 0.625 | 0.083 |

TABLE 2: PORTION OF THE ENTAILMENT-BASED VECTOR SPACE.

different from the standard treatment of polysemy in DSM, according to which all the different senses of a words are represented as a single vector obtained by abstracting over the different contexts of use of the different word senses (Arora *et al.* 2018; Boleda 2020).

### *3.3 Linear mapping*

Mappings between semantic spaces have been used in previous works for many purposes, among which the most relevant for our study are Făgărășan *et al.* (2015), Utsumi (2020), and Chersoni *et al.* (2021). Analogously to what has been done by these scholars, we learned a linear mapping between the distributional vectors and their representations in the entailment-based space and use these mappings to predict the vectors of a test group of untrained words. To generate the predicted DSM-based entailment vectors to evaluate against the target rating-based entailment vectors, we opted for a conventional ten-fold cross-validation. We thus split our dataset in 10 folds and generate predictions for each fold by learning a mapping from the remaining verbs. The coefficients of the mapping were estimated using the Partial Least Squared Regression method implemented in the scikit-lean Python library (Pedregosa *et al.* 2011), with the number of components[5] set to 10.

### *3.4 Evaluation*

We follow Chersoni *et al.* (2021) and Utsumi (2020) in evaluating our predicted vector by calculating the row-wise and the column-wide Spearman's $\rho$ between the rating-based entailment space and the DSM-based entailment space. Row-wise correlation measures the similarity of shape between the original and the inferred verb embeddings, while column-wise correlation measures the similarity of shape between the representation of the semantic proto-roles properties in the original space against those in the inferred space.

To assess the quality of the mappings generated from the different DSMs spaces, we repeated the procedure described in Section 3.3 on a matrix populated with values randomly sampled from a uniform distribution on the interval $[0,1]$ and treated the performance of this model as the baseline score.

## 4. RESULTS

In Figure 1 we use two metrics to summarize the performance of our DSMs. A first metric, the verb average correlation, is calculated by averaging over the row-wise correlations between the predicted entailment space and the space derived from the White *et al.* (2016)'s ratings. In intuitive terms, this metric tells us to what extent distributional information can be used to generate entailment-based information for novel verbs. As suggested by the left barplot, all models seem to perform better than the baselines model. Wilcoxon signed-rank tests with Holm-Bonferroni correction confirm this impression, in that all models other than `SVD.10w` ($W = 4309$, $p = .088$) are significantly better than `baseline` (all $W_s \leq 4172$, all $p_s < .05$).

In addition, by looking at this figure we can observe a marginal advantage of the neural network based DSMs (`SGNS`, `GloVe` and `fastText`) over the traditional `SVD` models, and a strong advantage of the syntactically typed context (`*.synt` models) over all the other types of contexts. However, the size of the differences between the DSM models are quite small, reaching statistical significance in a few sporadic cases. It should also be noted how the baseline score is rather high ($\rho = .626$), a fact that we tie to one of the key elements

[5] The optimal value of this hyper-parameter, defined as the number of components that maximizes the $R^2$, thus minimizing the mean square error, was inferred by running a 10-fold nested cross validation.

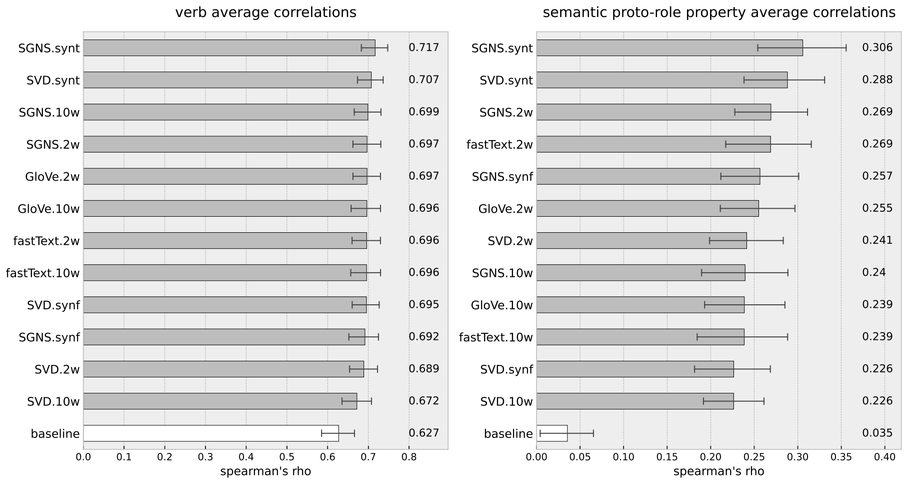

FIGURE 1: AVERAGE VERB (LEFT) AND PROTO-ROLE PROPERTY (RIGHT) CORRELATIONS FOR EACH TESTED DSM WITH 95% CONFIDENCE INTERVAL.

of the Dowtyan theory of semantic proto-roles: the association between grammatical function and semantic proto-roles. Subjects tend to have proto-agent properties, while object tend to have proto-patient properties. From this association, experimentally validated by works like Kako (2006b) and Reisinger *et al.* (2015), follows the fact that the vectors of our target entailment-based space are, to a certain extent, bound to share a similar structure in which some dimensions tend to be consistently scored higher than others. In our view, this partially explains the difference with baseline scores like those reported by Utsumi (2020) and Chersoni *et al.* (2021), whose target space is build from a dataset in which the ratings are more uniformly distributed.[6] Notwithstanding such a high baseline score, however, we see the significant improvement shown by virtually all DSMs as proving that some proto-role information is encoded by DSM models. The higher performance of the `*.synt` models suggests that syntactic information positively contributes to the ability of embeddings to encode properties pertaining to the verb argument structure, which are instead suboptimally captured by models based on bag-of-words co-occurrences.

The same conclusion can be reached by looking at the right plot in Figure 1, in which the correlation score of the baseline model is substantially null. The metric used in this plot is calculated by averaging over the column-wise corre-

[6] We also experimented with baselines analogue to those discussed by Utsumi (2020), recording values ranging from 0.572 to 0.644. We don't report these scores in details because they don't fit in the argumentative structure of the paper. Utsumi (2020)'s baselines, indeed, are model-specific and are difficult to interpret as the behavior of a property-agnostic model.

lations between the predicted entailment space and the space derived from the White *et al.* (2016)'s ratings. In intuitive terms, this metric can be understood as a measure of how our method is able to model the strength of association between verb arguments and proto-roles entailments. The lower average correlation scores suggest the higher complexity of this task. However, it should be noted that the overall ranking of the DSMs is similar to the one obtained in the case of the verb correlations. First, all models perform significantly better than the baseline model (all $W_s \leq 47$, all $p_s < .0001$). Moreover, there is further evidence of the advantage of the models relying on a syntactically typed context over all the other models: The correlation score of `SGNS.synt` is significantly higher than that of all other models (all $W_s \leq 168$, all $p_s < .05$), with the exception of the `SVD.synt` model ($W = 291$, $p = 1$), that in turn has a clear advantage over `SGNS.w10` and the subsequent models in the lower part of the plot (all $W_s \leq 171$, all $p_s < .05$).

Figure 2 shows the correlation coefficients for all the 41 typed properties we are focusing on in this study. This representation allows us to identify which properties are harder to model, and to understand whether this pattern is consistent across the argument positions. Overall, what immediately stands out from this plot is the main effect of the argument type and the interaction between argument type and entailment. As for the former, it is easy to spot how the `dobj` arguments are the easier to model, while in the `subj` arguments we see a clear distinction between reasonably well-modelled properties (`volition`, `sentient`, `awareness`, `volition`, `was for benefit`, `existed before` and, for some models, `change of state`) and properties that are problematic for all DSM models. Note how this contrast is even more accentuated in the top performing DSMs, i.e. `SGNS.synt` and `SVD.synt`.

A possible interpretation of this difficulty in the processing of `nsubj` arguments may be linked to the fact that the subject syntactic positions may realize different semantic roles across different clause structures [7]. As for the moment, we don't have the means for measuring the impact that the presence of syntactic alternations (and so the lacking of a one-to-one mapping between syntax and semantics) can affect the kind of semantic representation encoded by our static embeddings, in that this can be affected by many plausible factors (e.g. the language variety/varieties of the corpus, the type of documents in the corpus). This is, however, a issue that we planned to tackle in the future by extending our work to include contextualized DSMs like BERT (Devlin *et al.* 2019) and GPT (Radford *et al.* 2019; Brown *et al.* 2020).

[7] We'd like to thank Aaron Steven White for suggesting this possible interpretation in the review of an earlier version of this article.

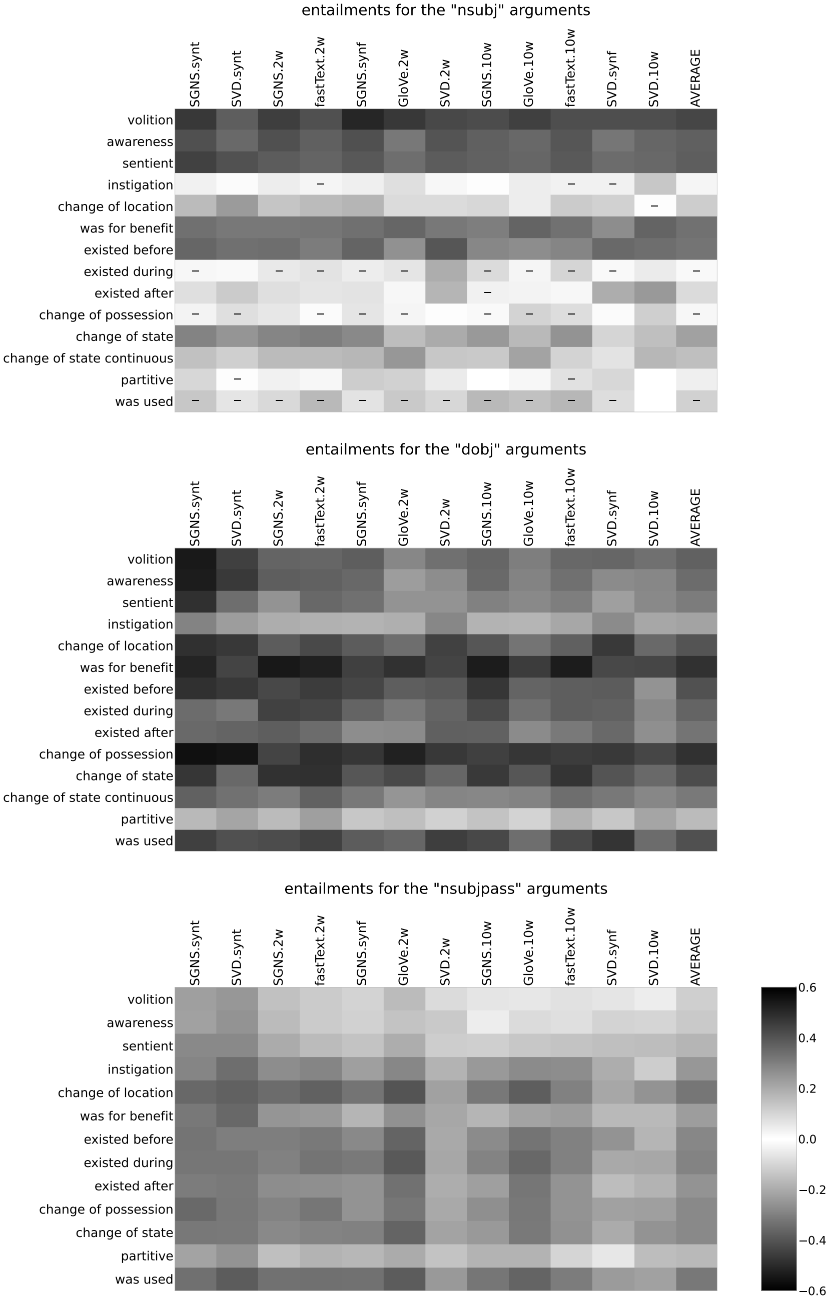

FIGURE 2: DETAILED PROTO-ROLE ENTAILMENT SPEARMAN'S CORRELATIONS FOR EACH TESTED DSM MODEL, ORGANIZED BY GRAMMATICAL RELATION. CELLS CONTAINING NEGATIVE VALUES ARE MARKED BY A MINUS SIGN.

While some properties appear to be overall genuinely difficult to model, as it seems to be the case for the `sentient` and the `partitive` entailments, the general trend is that the ranking of the properties by modeling difficulty is modulated by the argument type. Interestingly, this seems to be consistent with the Dowty (1991)'s lists of proto-role properties reported in (2) and (3) above. As for the `subj` arguments, the properties that are easier to model are `volition`, `sentient` and `awareness`, that can be easily mapped to the Dowtyan volitional involvement and sentence and/or perception proto-role properties. These entailments are strongly related to animacy, which is typically associoated to subjects. Another property that is easy to model for these arguments is the `change of state property`. A proto-agent property on which most DSM models seem to struggle is movement (i.e., the `change of location` entailment). All in all, this seems to suggest that our models are quite good at discriminating verbs taking strongly agentive subjects (e.g., `to kill`) from those occurring with "less agentive" subjects (e.g., `to affect`).

The pattern for the other two argument positions is rather different, with an advantage for those entailments that can be traced back to the Dowtyan proto-patient properties change of state and causally affected by another participant (e.g., `change of possession`, `change of state`, `change of location` and `was used`). Interestingly, the DSMs based on typed syntactic context are pretty good at modeling the `volition`, `awareness` and `sentient` properties also for the `dobj` arguments, in contrast with the other DSM models that seem to struggle with these relations.

## 5. UNACCUSATIVE VS. UNERGATIVE VERBS

To investigate the linguistic plausibility of the representation generated by the approach described in these pages, we decided to test it on a widely documented linguistic phenomenon: unaccusativity (Perlmutter 1978; Burzio 1986; Levin *et al.* 1995). According to the Unaccusative Hypothesis first introduced by Perlmutter (1978) and elaborated by Burzio (1986), the only argument of intransitive verbs like those in (4) is an underlying object, even if it is superficially realized as a subject.

(4) *John aged / appeared / grew / died*

Due to this behaviour, it is common practice to label such verbs *unaccusative* (or *ergative*) verbs and to contrast them with the complementary class of intransitive verbs: the *unergatives*, sometimes also labelled *pure intransitive*.

Unergative verbs like those in (5) are intransitive verbs whose surface subject is an actual underlying subject.

(5) *John slept / jog / worked*

From a semantic perspective, the opposition between unaccusative and unergative verbs has been tied to the different semantic roles of the subject: AGENT in the case of the unergative verbs, THEME or PATIENT in the case of the unaccusative verbs (e.g., Pustejovsky & Batiukova 2019: 225). Roughly speaking, then, the key difference between the verbs in (4) and those in (5) is that in the former *John* is affected by the event, while in the latter he is the person performing the action (i.e., the "doer"). Dowty (1991: 605-613) characterized the unaccusative/unergative dichotomy in terms of proto-roles: Unaccusative verbs are intransitive verbs that entail PROTO-PATIENT properties on their subject, while unergative verbs entail PROTO-AGENT properties on their subject. In the events described by the examples in (5), the subject is performing an action depending on his own volition, thus displaying the volition involvement and causing event PROTO-AGENT properties. Conversely, in the events described in (4), the subject undergoes a change of state that may be caused by another participant, thus displaying two PROTO-PATIENT properties.

From this quick characterization of the unergative/unaccusative dichotomy it is easy to see how the modelling unaccusativity could be an appealing testing ground for our approach. Given that unaccusativity is a phenomenon that lays at the syntax-semantic interface, we wonder whether a entailment space inferred from a DSM is able to tease unaccusatives and unergatives apart, and what kind of proto-role entailment representation it associates with these two classes. We thus set a small experiment in which we learned a mapping from what has been the best performing DSM in Section 4 (i.e., `SGSN.synt`) to an entailment space encoding only the ratings from the White *et al.* (2016)'s dataset that describe a `nsubj` arguments. We decided to filter out the ratings of the `dobj` and `nsubjpass` arguments because our goal is to derive the property of the subject of an intransitive (unergative or unaccusative) verb. Again, we used a Partial Least Squared Regression to learn the mapping between the DSM and the entailment space, setting the number of components to 5 due to the reduced dimensionality of our target representation. We trained the mapping for all the 149 unambiguously transitive verbs used in our main experiment with the exception of few of our target verbs. We then applied this mapping to infer an entailment space for the `nsubj` arguments of the following intransitive verbs[8]:

[8] The verbs of this test set were chosen by browsing the available literature on the topic as well as

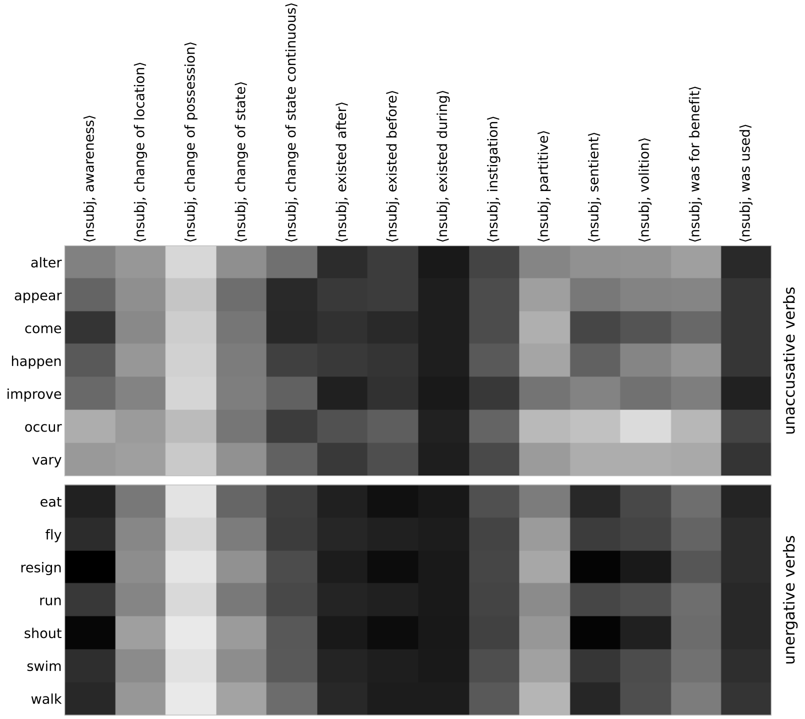

FIGURE 3: INFERRED ENTAILMENT-BASED REPRESENTATION FOR OUR TEST SET OF INTRANSITIVE VERBS. DARKER HUES DENOTE HIGHER VALUES.

- Unaccusative verbs: `alter`, `appear`, `come`, `happen`, `improve`, `occur`, `vary`
- Unergative verbs: `eat`, `fly`, `resign`, `run`, `shout`, `swim`, `walk`

The heatmaps in Figure 3 visually represent, for each unaccusative (top) and unergative (bottom) verb, the inferred rating for each proto-role entailment. Darker hues denote higher values, while white cells indicate a null rating. For instance, by looking at the first and at the third cell of the second third row of the bottom heatmap we can conclude that the subjects of the verb `to resign` is highly aware but does not undergo a change of possession.

Even a visual inspection of the two plots in Figure 3 reveals that the inferred representations associated to the unergative and to the unaccusative verbs are quite different. Moreover, this contrast partly mirrors the linguistic

a few English grammar books and by randomly sampling over the set of frequent ($f < 1,000$) lemmas that whose occurrences in our corpus were tagged as a verb more than 50% of the times.

characterization of unaccusativity: the subject of the unergative verbs are indeed more strongly associated with PROTO-AGENT qualities (mainly `volition`, `sentient`, `awareness` and `was for benefit`) than those of the unaccusative verbs. Mann-Whitney U tests confirm that these differences are statistically significant ($U \leq 5$, $p < .01$). Other qualities that are significantly more associated with the subject of unergative verbs are `existed before`, `existed during` and `existed after` ($U \leq 8$, $p < .05$), three properties that, coherently with a Dowtyian perspective, were judged as agentive by the the annotators from Reisinger *et al.* (2015). It should also be noted how the fact that the subjects of some unaccusative verbs possess some low-to-mild agentive quality is consistent with the fact that some of these verbs (those expressing a change of state) can be used transitively by participating in the so-called causative alternation (Levin *et al.* 1995).

(6) a. *Prices altered significantly in the aftermath of the COVID crisis*
b. *I haven't altered my look in the last ten years*

If the agentive properties appear to be represented in a way that mirror the unaccusative/unergative dichotomy, the same cannot be said of their patientive counterparts. The only PROTO-PATIENT quality strongly possessed by the subjects of the unaccusative verbs is the `change of possession` one ($U = 0$, $p < .01$). All the other differences are not statistically significant ($U \geq 8$, $p > .05$), not even the differences in that `change of state` property that the linguistic literature has shown to be crucial in the determination of the unaccusative status of a verb (see e.g. Levin *et al.* 1995).

Notwithstanding this limitation, it is undeniable how our strategy associated different representations to the two major types of intransitive verbs, and and we are comfortable in concluding that it seems to be able to properly represent the non-agentive nature of the subject of the unaccusative verbs.

## 6. TRANSITIVE PSYCH-VERBS

Another opposition that has been characterized in terms of semantic roles is that between the two main classes of transitive verbs expressing psychological states (a.k.a. transitive *psych-verbs*), which following Levin (1993: 188-193) we will label as *amuse* verbs and *admire* verbs. These two classes are usually taken apart by the syntactic realization of the EXPERIENCER:[9] the subject for the admire verbs; the object for the amuse verbs.

[9] The EXPERIENCER role can be defined as the entity psychologically affected by the event described by the verb (Pustejovsky & Batiukova 2019: 29).

Following Levin (1993: 191), we can characterize amuse verbs as the subclass of psych-verbs that describe a change in psychological or emotional state, as exemplified in the following sentence:

(7) *I/My jokes amuse only her*

In this example the object (i.e., *her*) denotes the person that is emotionally affected by the event, while (i.e., *My jokes* or *I*) the subject denotes the entity causing the psychological change. On the other hand, the subjects of admire verbs are better described as EXPERIENCERS rather than AGENTS, as shown by the example in (8). The object roles of these verbs are described as THEME or STIMULUS:

(8) *A few tourists admired the Church of San Pantalon*

Analogously to what we've done in Chapter 5, here again we ask ourselves whether the entailment space inferred from a DSM is able to tease admire and amuse verbs apart, and what kind of proto-role properties it associates with these two classes. We than performed a small experiment in which we learn a mapping from `SGNS.synt` to the space build from the White *et al.* (2016)'s ratings. Given that we work with transitive verbs in this case, we departed from the unaccusative/unergative experiment by building a space encoding all the entailment relations for all the possible grammatical relations. Again, we chose Partial Least Squared Regression as our learning algorithm but we set the number of components to 10. We trained the mapping for all the 152 verbs we have used in our main experiment in Section 4 and applied this mapping to infer an entailment space for a set of verbs of interest. Our target lemmas were randomly selected from the verbs listed in the classes 31.1 (amuse verbs) and 31.2 (admire verbs) of the Levin (1993)'s classification, from which we removed all the verbs that were listed also in the *marvel* class (i.e., that could be also used intransitively) and all the lemmas that were either infrequent or grammatically ambiguous. We ended up with the following list of 28 verbs:

- Admire verbs: aggravate, agitate, amuse, captivate, confound, confuse, disturb, encourage, haunt, inspire, intimidate, pacify, placate, reassure
- Amuse verbs: admire, adore, appreciate, cherish, deplore, despise, detest, distrust, enjoy, exalt, mourn, resent, stand, tolerate

The heat-maps in Figure 4 depict, for each admire (top) and amuse (bottom) verb, the inferred score for each $\langle grammatical\_function, property \rangle$ pair. Even at first sight it is evident that our mapping neatly dissociated the admire and the amuse verbs. Some of the differences follows from the different roles of the arguments of these two classes of verbs, some other – we speculate – may follow from well studied linguistic properties of these two verbal classes, some others are instead harder to interpret.

An eye-popping contrast in these representations is the one involving the `sentience` entailment of the `obj` arguments. The objects of the admire verbs are significantly less sentient than their amuse counterparts, as confirmed by a Mann-Whitney U test ($U = 10$, $p < .0001$). In a similar fashion, we can see that the `change of state` and `change of possession` property on the `obj` arguments are significantly higher in the amuse verbs (all $U_s \leq 28$, all $p_s < .001$). Both these observations are consistent with Dowty (1991: 579-580)'s analysis of the objects of experiencer-object verbs as bearing the PROTO-AGENT property of sentience and the PROTO-PATIENT property of undergoing a change of state.

Another big difference in the patterns of entailment of these two classes concerns the agentivity of the subject. The subjects of the amuse verbs, indeed, show significant lower scores than their admire counterparts in all the main markers of agentivity: `awareness`, `sentient` and `volition` (all $U_s \leq 30$, all $p_s < .001$). An exception to this pattern regards the `instigation` property, in which the advantage of the admire verbs is not statistically significant ($U = 82$, $p \approx .24$). In our view, this phenomenon can be traced back to the fact, discussed by Grimshaw (1990), that the subject of some amuse verbs can receive an agentive interpretation, as is the case for the (7) case above, while verbs like *concern* do not, as shown in the following example:

(9) *Your happiness concerns me*

Finally, the reader may notice an evident difference in the scores associated with the `nsubjpass` arguments of these two classes. We do not feel comfortable interpreting this difference, however, due to the data sparsity issue that afflicts both the distributional space and the rating-based entailment space. As for the former, indeed, it should be considered that the `nsubjpass` ratings constitute a small minority of the overall ratings, amounting to approximately 5% of the selected judgments. As for the DSM, it should be stressed that all the models we considered are built from an unlemmatized corpus, so that the verbal forms that are typically used in passive context typically receive a different representation than the one we have used as the reference form in the experiments reported in these pages.

Notwithstanding the latter limitation, we again see the reported results as demonstrating that the proto-role entailment representation we have inferred

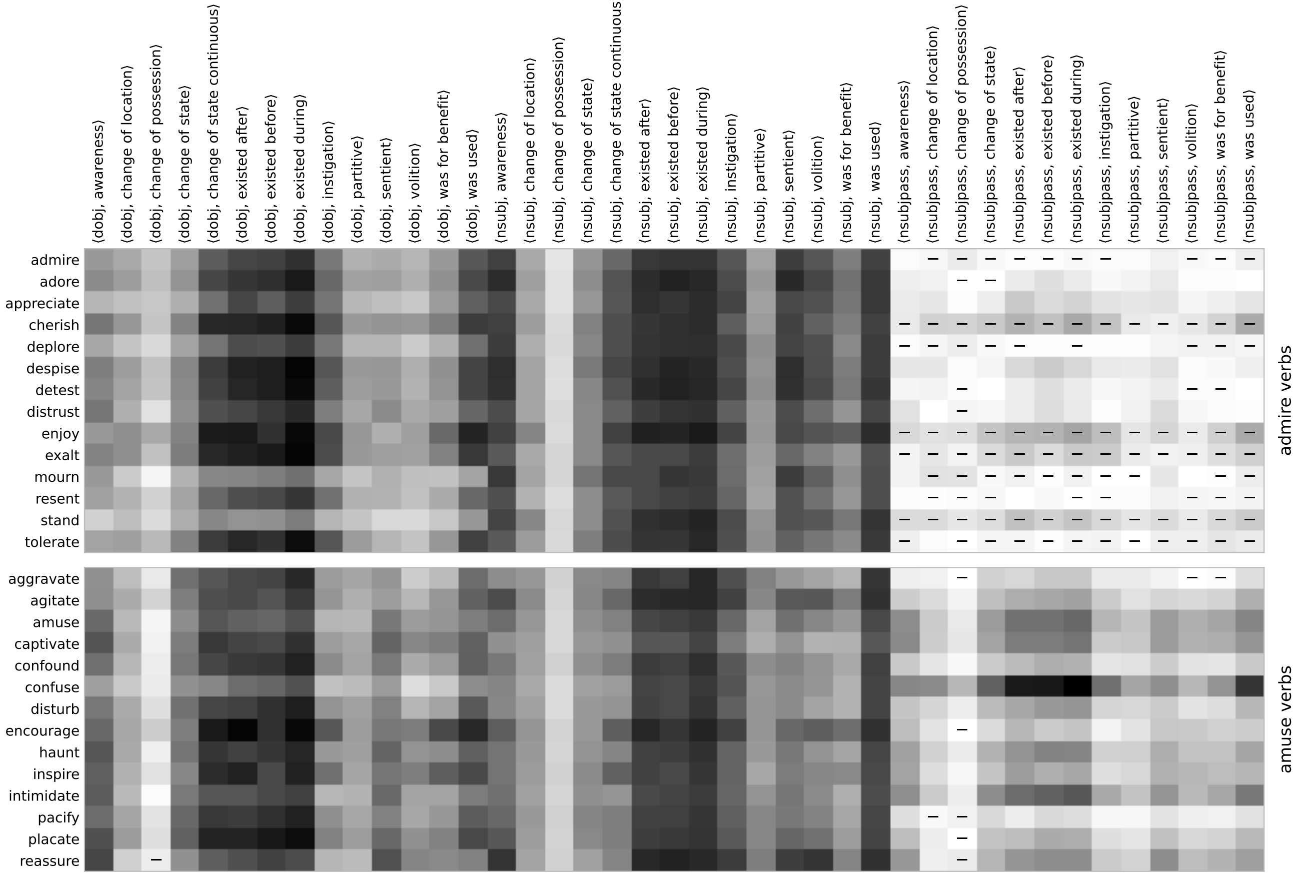

FIGURE 4: INFERRED ENTAILMENT-BASED REPRESENTATION FOR OUR TEST SET OF TRANSITIVE PSYCH-VERBS. DARKER HUES DENOTE HIGHER VALUES. CELLS CONTAINING NEGATIVE VALUES ARE MARKED BY A MINUS SIGN.

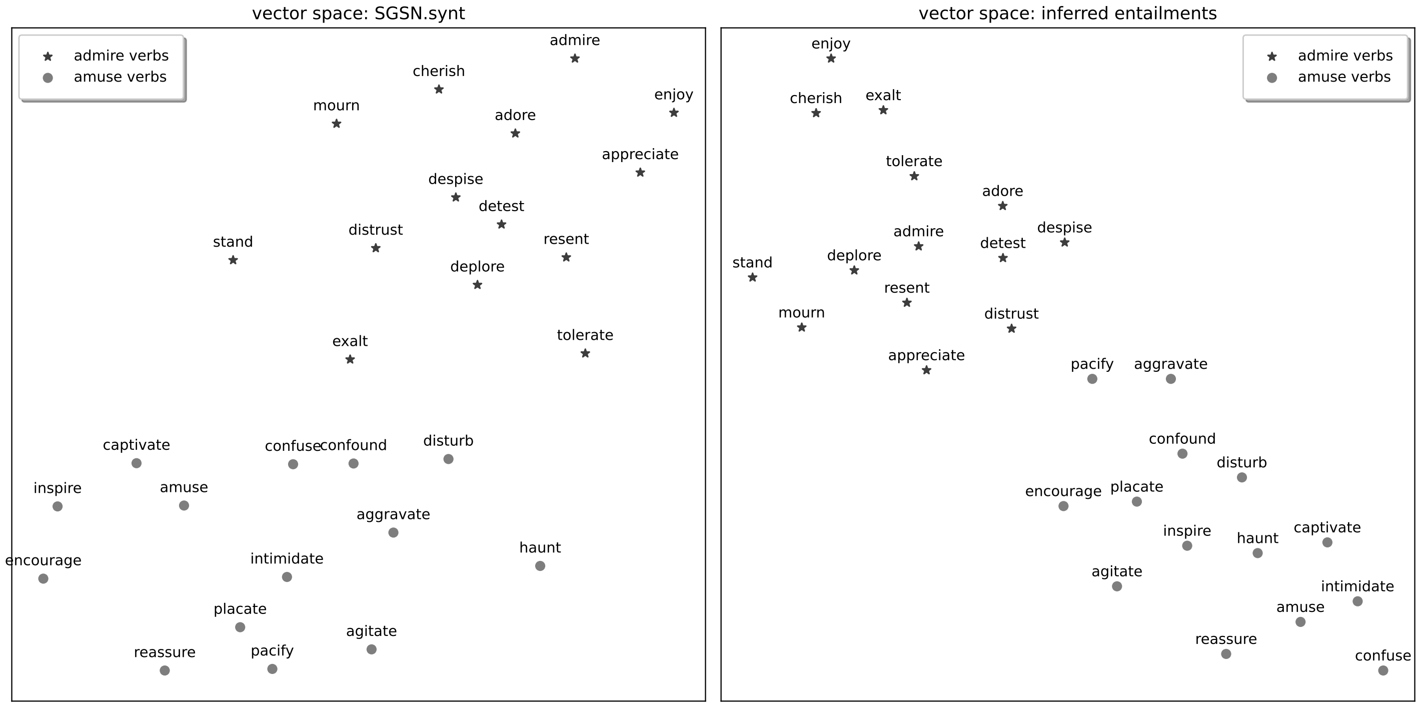

FIGURE 5: TWO-DIMENSIONAL REPRESENTATION OF OUR TARGET PSYCH-VERBS IN THE ORIGINAL DMS (LEFT) AND IN THE INFERRED SPACE (RIGHT). DIMENSIONALITY REDUCTION OBTAINED USING T-SNE (PERPLEXITY: 5; STEP: 1,000).

from our source DSM is able to disentangle the two verb classes along some dimensions that are interpretable and, more importantly, linguistically plausible. The right plot of Figure 5 represents our verbs in a bi-dimensional space built by using the t-distributed Stochastic Neighbor Embedding (t-SNE: van der Maaten & Hinton 2008) dimensionality reduction technique. It shows a neat separation between the admire and the amuse verbs, as a further proof of the fact that the representations we have inferred for these two classes are actually different. However, it should be noted that a similar pattern can be drawn by relying of the DSM alone, as shown by the left plot of the same figure. This illustrates a crucial aspect of the method we have borrowed from Mikolov *et al.* (2013b), Făgărăşan *et al.* (2015), Utsumi (2020), and Chersoni *et al.* (2021): It does not create new information, but it simply extracts information encoded in the verb embeddings. The entailment-based information that we commented in our experiments, indeed, are already in the vector spaces built simply by looking at the way words are used, and the mapping distills it in order to make it human-interpretable.

## 7. CONCLUDING REMARKS

In this paper, we reported an experiment in which we inferred the semantic content of Dowty (1989, 1991)'s proto-roles from the distributional information encoded in a variety of vector spaces built by using a restricted group of static DSMs. To the best of our knowledge, this is the first time embedding-to-feature mappings are applied to the case of verb argument structure and semantic role entailments. We evaluated this approach by comparing the inferred representation against the proto-roles rating collected by White *et al.* (2016). We moreover conducted a qualitative analysis focusing on the inferred representations associated with verb classes oppositions that are often characterized in terms of thematic roles: that between unaccusative and unergative verbs and the opposition between the two main classes of psych-verbs. All in all, we interpret the results of our experiments as showing that our method has potential use not only in linguistics, but even for practical applications such as the Semantic Proto-Role Labelling task proposed by Van Durme, Rawlins and colleagues (Reisinger *et al.* 2015; Teichert *et al.* 2017; White *et al.* 2017; Rudinger *et al.* 2018) and the related Universal Decompositional Semantic Parsing task tackled by Stengel-Eskin *et al.* (2020, 2021).

The main contributions of our work, however, is not the technique itself, but the demonstration that proto-roles information can be inferred from distributional embeddings, to an extent that should be matter of further research. As such, then, this work enlarges the literature trying to characterize what kind of semantic information in encoded in word embeddings.

Our works and our conclusions, however, have some limitations. An important limitation has to do with the variety of DSM techniques tested. All of our spaces, indeed, are built from spaces that are not lemmatized nor PoS-tagged. As a consequence, our spaces have different entries for different inflected forms of the same verb (e.g., two different vectors for `play` and `played`), as well as a unique entry for lemmas belonging to different parts of speech (e.g., a unique vector for the verb `to mail` and for the noun `noun`). These two sides of the same coin affected our experiment in many ways, among which the fact that they forced us to put in place filtering strategies that dramatically limited the inventory of train and test lexical entries. Moreover, it is not unreasonable to hypothesize that the linguistic-agnosticism of our DSMs may had an effect on our inferred entailment representations by providing scarce and/or conflicting evidence. This is consistent with the observation that the best performing models are the syntactic typed spaces. In order to generalize our findings, it is vital to evaluate a much larger set of verbs and to be able to use as much linguistic information as possible, and this will be possible only by working with a DSM build on a pre-processed corpus.

However, DSMs are not the only source of sparsity in our experiment. In the ratings collected by White *et al.* (2016), indeed, not all the verb nor all the $\langle grammatical function, property \rangle$ pairs received the same number of judgments. This contributed to the limited the inventory of train verbs in our experiment and arguably even influenced part of our results. Surely, the sparsity of the `iobj` arguments in the corpus did not allow us to study the semantic properties of this grammatical role, but it is not unreasonable to think that sparsity may had an effect on the representation of the `nsubjpassive` properties as well. Driven by the promising results of this study, we are planning a further collection of proto-roles judgments built on the work by Reisinger *et al.* (2015) and White *et al.* (2016).

Finally, in this work we did not experiment with contextualized DSMs like BERT (Devlin *et al.* 2019) and GPT (Radford *et al.* 2019; Brown *et al.* 2020), and we tested only one mapping strategy, thus ignoring other machine learning approaches like the Multi Layer Perceptron tested by Utsumi (2020). Even if we are currently planning to pursue our research by filling this void, we do not see these as real limitations of our work, in that we doubt that our choice to focus only on static DSM and to use only a linear mapping may have influence the take-home-message of this paper: somewhere in each verb embedding there is some proto-role information learned simply by looking at the company the verbs keep.

# REFERENCES

Anderson, A.J., D. Kiela, S. Clark & M. Poesio (2017). Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5. 17–30. doi:10.1162/tacl_a_00043.

Arora, S., Y. Li, Y. Liang, T. Ma & A. Risteski (2018). Linear Algebraic Structure of Word Senses, with Applications to Polysemy. *Transactions of the Association for Computational Linguistics*, 6. 483–495. doi:10.1162/tacl_a_00034.

Baroni, M. & A. Lenci (2010). Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4). 673–721. doi:10.1162/coli_a_00016.

Baroni, M., S. Bernardini, A. Ferraresi & E. Zanchetta (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3). 209–226. doi:10.1007/s10579-009-9081-4.

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. doi:10.1080/02643294.2016.1147426.

Bojanowski, P., E. Grave, A. Joulin & T. Mikolov (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5. 135–146. doi:10.1162/tacl_a_00051.

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*, 6(1). 213–234. doi:10.1146/annurev-linguistics-011619-030303.

Bommasani, R., K. Davis & C. Cardie (2020). Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online, 4758–4781.

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.

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.

Caliskan, A., J.J. Bryson & A. Narayanan (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334). 183–186. doi:10.1126/science.aal4230.

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

de Marneffe, M.C., C.D. Manning, J. Nivre & D. Zeman (2021). Universal Dependencies. *Computational Linguistics*. doi:10.1162/coli_a_00402.

Devereux, B.J., L.K. Tyler, J. Geertzen & B. Randall (2014). The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, 46(4). 1119–1127. doi:10.3758/s13428-013-0420-4.

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.

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.

Dowty, D.R. (1991). Thematic Proto-roles and Argument Selection. *Language*, 67. 547–619. doi:10.2307/415037.

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. doi:10.1006/jmla.2000.2728.

Făgărăşan, 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.

Grand, G., I.A. Blank, F. Pereira & E. Fedorenko (2018). Semantic projection: recovering human knowledge of multiple, distinct object features from word embeddings. *arXiv preprint arXiv:1802.01241*.

Grimshaw, J. (1990). *Argument Structure*. Cambridge, MA: The MIT press.

Hare, M., M. Jones, C. Thomson, S. Kelly & K. McRae (2009). Activating event knowledge. *Cognition*, 111(2). 151–167. doi:10.1016/j.cognition.2009.01.009.

Harris, Z.S. (1954). Distributional Structure. *Word*, 10(2-3). 146–162.

Huth, A.G., W.A. De Heer, T.L. Griffiths, F.E. Theunissen & J.L. Gallant (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600). 453–458. doi:10.1038/nature17637.

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

Jain, M., P. Goel, P. Singla & R. Tehlan (2021). Comparison of Various Word Embeddings for Hate-Speech Detection. In A. Khanna, D. Gupta, Z. Pólkowski, S. Bhattacharyya & O. Castillo (eds.) *Data Analytics and Management. Lecture Notes on Data Engineering and Communications Technologies*. Singapore: Springer, 251–265.

Johns, B.T. & M.N. Jones (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, 4(1). 103–120. doi:10.1111/j.1756-8765.2011.01176.x.

Jones, M.N., T.M. Gruenenfelder & G. Recchia (2018). In defense of spatial models of semantic representation. *New Ideas in Psychology*, 50. 54–60. doi:10.1016/j.newideapsych.2017.08.001.

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

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

Kumar, P.S., R.B. Yadav & S.V. Dhavale (2021). A comparison of pre-trained word embeddings for sentiment analysis using Deep Learning. In *International Conference on Innovative Computing and Communications*. Springer, 525–537.

Landauer, T.K. & S.T. Dumais (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2). 211–240. doi:10.1037/0033-295x.104.2.211.

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.

Lebani, G.E., A. Bondielli & A. Lenci (2015). You Are What you Do. An Empirical Characterization of the Semantic Content of the Thematic Roles for a Group of Italian Verbs. *Journal of Cognitive Science*, 16(4). 401–430. doi:10.17791/jcs.2015.16.4.401.

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

Lenci, A., G.E. Lebani & L.C. Passaro (2018). The emotions of abstract words: A distributional semantic analysis. *Topics in cognitive science*, 10(3). 550–572. doi:10.1111/tops.12335.

Lenci, A., M. Sahlgren, P. Jeuniaux, A.C. Gyllensten & M. Miliani (2021). A comprehensive comparative evaluation and analysis of Distributional Semantic Models. *arXiv preprint arXiv:2105.09825*.

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

Levin, B. & M. Rappaport Hovav (2005). *Argument Realization*. Research Surveys in Linguistics, Cambridge, UK: Cambridge University Press.

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

Levy, O. & Y. Goldberg (2014). Dependency-Based Word Embeddings. In *Proceedings of the 52$^{nd}$ Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland, 302–308.

Levy, O., Y. Goldberg & I. Dagan (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3. 211–225. doi:10.1162/tacl_a_00134.

Loureiro, D., K. Rezaee, M.T. Pilehvar & J. Camacho-Collados (2021). Analysis and Evaluation of Language Models for Word Sense Disambiguation. *Computational Linguistics*, 28. doi:10.1162/coli_a_00405.

Lund, K. & C. Burgess (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2). 203–208. doi:10.3758/BF03204766.

Mandera, P., E. Keuleers & M. Brysbaert (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92. 57–78. doi:10.1016/j.jml.2016.04.001.

Manning, C.D., M. Surdeanu, J. Bauer, J.R. Finkel, S. Bethard & D. McClosky (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland, 55–60.

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

McRae, K., G.S. Cree, M.S. Seidenberg & C. McNorgan (2005a). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4). 547–559. doi:10.3758/BF03192726.

McRae, K., M. Hare, J.L. Elman & T.R. Ferretti (2005b). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, 33(7). 1174–1184. doi: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*.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado & J. Dean (2013c). Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. 3111–3119.

Mitchell, T.M., S.V. Shinkareva, A. Carlson, K.M. Chang, V.L. Malave, R.A. Mason & M.A. Just (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880). 1191–1195. doi:10.1126/science.1152876.

Padó, S. & M. Lapata (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2). 161–199. doi:10.1162/coli.2007.33.2.161.

Palmer, M., D. Gildea & P. Kingsbury (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1). 71–106. doi:10.1162/0891201053630264.

Passaro, L.C., A. Bondielli & A. Lenci (2017). Learning affect with distributional semantic models. *Italian Journal of Computational Linguistics*, 3(3-2). 23–36. doi:10.4000/ijcol.550.

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. doi:10.5555/1953048.2078195.

Pennington, J., R. Socher & C. Manning (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, 1532–1543.

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

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

Qi, Y., D.S. Sachan, M. Felix, S.J. Padmanabhan & G. Neubig (2018). When and why are pre-trained word embeddings useful for neural machine translation? *arXiv preprint arXiv:1804.06323*.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei & I. Sutskever (2019). Language Models are Unsupervised Multitask Learners. *OpenAI blog*.

Recchia, G. & M.M. Louwerse (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *Quarterly journal of experimental psychology*, 68(8). 1584–1598. doi:10.1080/17470218.2014.941296.

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. doi:10.1162/tacl_a_00152.

Riordan, B. & M.N. Jones (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2). 303–345. doi:10.1111/j.1756-8765.2010.01111.x.

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

Rubinstein, D., E. Levi, R. Schwartz & A. Rappoport (2015). How Well Do Distributional Models Capture Different Types of Semantic Knowledge? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, 726–730.

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.

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.

Sommerauer, P. & A. Fokkens (2018). Firearms and Tigers are Dangerous, Kitchen Knives and Zebras are Not: Testing whether Word Embeddings Can Tell. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium, 276–286.

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.

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. doi:10.1162/tacl_a_00396.

Teichert, A., A. Poliak, B.V. Durme & M. Gormley (2017). Semantic Proto-Role Labeling. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 4459–4465.

Utsumi, A. (2015). A complex network approach to distributional semantic models. *PloS one*, 10(8). e0136277. doi:10.1371/journal.pone.0136277.

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

van der Maaten, L. & G. Hinton (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86). 2579–2605.

Vinson, D.P. & G. Vigliocco (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40(1). 183–190. doi:10.3758/BRM.40.1.183.

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.

White, A.S., K. Rawlins & B. Van Durme (2017). The Semantic Proto-Role Linking Model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain, 92–98.

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.

Yu, L.C., J. Wang, K.R. Lai & X. Zhang (2017). Refining Word Embeddings for Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark, 534–539.

*Gianluca E. Lebani*
Ca' Foscari University of Venice
Ca' Bembo, Fondamenta Tofetti, Dorsoduro 1075 - 30123 Venice
Italy
e-mail: `gianluca.lebani@unive.it`

*Alessandro Lenci*
University of Pisa
via Santa Maria 36 - 56126 Pisa
Italy
e-mail: `alessandro.lenci@unipi.it`