Distributional models of category concepts based on names of category members

Matthijs Westera[*], Abhijeet Gupta[+], Gemma Boleda[‡,§], and Sebastian Padó[†]

[*]Leiden University Centre for Linguistics, Universiteit Leiden, Leiden, 2311 BE, The Netherlands

[+]Institut für Sprache und Information, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, 40225, Germany

[‡]Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona, 08018, Spain

[§]ICREA, Barcelona, 08010, Spain

[†]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Stuttgart, 70569, Germany

August 30, 2021

Abstract

Cognitive scientists have long used distributional semantic representations of categories. The predominant approach uses distributional representations of category-denoting nouns, like "city" for the category city. We propose a novel scheme that represents categories as prototypes over representations of names of its members, such as "Barcelona", "Mumbai", and "Wuhan" for the category city. This name-based representation empirically outperforms the noun-based representation on two experiments (modelling human judgments of category relatedness and predicting category membership) with particular improvements for ambiguous nouns. We discuss the model complexity of both classes of models and argue that the name-based model has superior explanatory potential with regard to concept acquisition.

Keywords: concepts, distributional semantics, prototype theory, ambiguity, relatedness

1 Introduction

Categories are fundamental to our perception and understanding of the world (Gärdenfors, 2000; Murphy, 2002; Rosch, 1975; E. Smith & Medin, 1981). For instance, if we recognize an entity as a member of the category of scientists, then we can infer that this entity is a human and does research. These inferences showcase the richness of our mental concepts of categories, including properties that enable member recognition as well as inter-category relationships such as relatedness or subsumption. This richness is partially mirrored in language, chiefly in the structure of our mental lexicon (Murphy, 2002) and in the ways in which words are used (Harris, 1954). The latter observation has led to a rich literature

Correspondence should be sent to the first author: Matthijs Westera, Reuvensplaats 3-4, 2311 BE Leiden, The Netherlands. E-mail: m.westera@hum.leidenuniv.nl

on computational models based on the so-called distributional hypothesis (Baroni, Dinu, & Kruszewski, 2014; Landauer & Dumais, 1997; G. A. Miller & Charles, 1991) (see Section 2 for details).

A common simplification both on the cognitive and on the computational side of distributional semantics research is the use of words as stand-ins for categories. In categorization research, many studies collect experimental data on category-denoting words (e.g., "scientist"), which is then interpreted in terms of category concepts, since "A function of word meaning, much studied in psychology, is to categorize the world into labelled classes." (Hampton, 2015). Analogously, in computational modeling, category-denoting nouns are often used as proxies for categories; for instance, the distributional representation of "scientist" is used to represent the concept of the category of scientist (e.g., work on semantic relations such as Baroni and Lenci (2011)).

In this paper, we focus on the use of distributional models to represent categories, and on how different kinds of linguistic data (common nouns and entity names) affect model quality. Here, it is crucial to note that, despite the successes of distributional models (see, e.g., Lenci 2008; G. A. Miller and Charles 1991), the distributional representation of a word is at best only a proxy for the category concept this word is typically used to express. The quality of this proxy depends on the degree of alignment between words and the concepts they are used to model (see Louwerse 2008, 2018). Polysemy is a particularly pervasive cause of misalignment (Cruse, 1986; Murphy, 2002), e.g., the word vector for "mouse" will blend speakers' uses of that word to refer to animals and to computer devices, among others. Other factors include figures of speech such as metaphor (e.g., calling a smart kid "scientist"), metonymy, and connotational aspects such as such as register (Fellbaum, 1998; Jackendoff, 1990; G. A. Miller & Charles, 1991; Murphy, 2002). These factors influence a word's distribution in ways that are orthogonal to the category it is typically used to denote.

Importantly, not all parts of the lexicon are equally subject to the aforementioned factors, and the adequacy with which distributional representations model extra-linguistic concepts can be expected to vary accordingly. Compared to common nouns, names (proper nouns) such as "Marie Skłodowska Curie" and "Albert Einstein" are 'rigid' (Kripke, 1980) in that they are almost universally used to refer to particular individuals (here, the scientists Marie Skłodowska Curie and Albert Einstein, respectively). Indeed, in computational linguistics, names have been found to exhibit lower polysemy than nouns, higher inter-annotator agreement in manual disambiguation, and lower error rates in automatic disambiguation (Chang, Spitkovsky, Manning, & Agirre, 2016). As a consequence, distributions of names are more uniform than those of nouns, which could result in more 'focused' distributional representations, with the potential to provide a superior ingredient for category representations.

Based on this motivation, our main aim in this paper is to investigate the simple but hitherto unexplored hypothesis that names of members of a category (e.g., for scientist, "Albert Einstein" and "Marie Skłodowska Curie") provide a better distributional representation of category concepts than the corresponding category-denoting nouns. This approach builds on existing work which shows that name vectors have been found to be reasonable proxies for entity representations (Gupta, Boleda, Baroni, & Padó, 2015; Herbelot, 2015; Toutanova et al., 2015). Furthermore, name vectors can be aggregated into category representations using cognitively plausible mechanisms like prototype formation by averaging (Rosch, 1975). Indeed, we consider our name-based approach to be a basic, distributional implementation of prototype theory (we consider also an analogous approach based on exemplar theory (Nosofsky, 1986)). More generally, in deriving category concepts from (a linguistic proxy for) entities, our name-based approach aligns more closely than the standard, noun-based approach with the literature on concept acquisition, which predominantly concentrates on the acquisition of categories from exemplars (e.g., Xu 2002).

To assess the advantages and disadvantages of name- vs. noun-based models empirically, we compare them in two experiments that model human judgments of category relatedness and category membership

(Section 4), both showing substantial superiority of the name-based model. We also compare the two models on a more conceptual level (Section 5), by showing that the noun-based model indeed suffers from polysemy, by assessing the relative complexity of each model, and by exploring their explanatory potential from the aforementioned perspective of concept acquisition. Note that this work concerns only the potential of names vs. nouns for representing extra-linguistic category concepts, not for representing the lexical meanings of nouns (where the noun-based model is expected to do better; see Section 6). All data and analysis code for the current paper is available at `https://osf.io/txchw/`.

2 Background

2.1 Foundations and Applications of Distributional Semantics

Distributional semantics is rooted in the hypothesis that words with similar meanings have similar linguistic distributions (Harris, 1954). In such models, word meaning is typically modeled in terms of high-dimensional vectors, also known as embeddings in the neural network literature, with values representing abstractions of word usage in linguistic sources such as textual corpora. Vectors of words with similar distributions in the corpus, such as "physicist" and "scientist", end up close together in this space. Early work computed vectors on the basis of word-context occurrence counts (e.g., Latent Semantic Analysis, Landauer and Dumais 1997). Current standard procedure is to train neural networks on word-context pairs given some predictive task, such as predicting the word given its linguistic context or vice versa (e.g., Word2Vec, Mikolov, Sutskever, Chen, Corrado, and Dean 2013). This outperforms traditional approaches on a variety of tasks (Baroni et al., 2014; Levy, Goldberg, & Dagan, 2015). Latest developments involve neural models which provide contextualized representations of words (Devlin, Chang, Lee, & Toutanova, 2019; Peters et al., 2018), i.e., the same word is assigned a different vector depending on the context in which it occurs. We relate these representations to our approach in Section 6.

Distributional semantic models are ubiquitous in Cognitive Science and Artificial Intelligence to model word meaning and conceptual knowledge. Examples of phenomena where distributional models have shown their usefulness are semantic priming (Griffiths, Steyvers, & Tenenbaum, 2007; Landauer & Dumais, 1997), word relatedness (Agirre et al., 2009), hypernymy (Levy, Remus, Biemann, & Dagan, 2015; Roller, Erk, & Boleda, 2014), property prediction (Baroni, Murphy, Barbu, & Poesio, 2010; Făgărăşăn, Vecchi, & Clark, 2015), analogy (Turney & Litman, 2005), metonymy (Shutova, Kaplan, Teufel, & Korhonen, 2013), and multimodal knowledge (Bruni, Boleda, Baroni, & Tran, 2012). Some studies establish similarities between distributional semantic representations and neuro-imaging data (Huth, De Heer, Griffiths, Theunissen, & Gallant, 2016; T. M. Mitchell et al., 2008; Søgaard, 2016). Finally, distributional semantics has been argued to represent not just a useful tool but also a cognitively plausible model of acquisition (e.g., Günther, Rinaldi, and Marelli 2019; Jones, Hills, and Todd 2015; Mandera, Keuleers, and Brysbaert 2017). In Section 5 we take up this point and discuss the cognitive adequacy of the name-based and noun-based models.

2.2 Categories and Entities in Distributional Semantics and Cognitive Science

Distributional semantic models are successful on phenomena which concern not just the words themselves, but the extra-linguistic, real-world categories to which they can refer – for instance, "dog" is a hyponym of "animal" by virtue of actual dogs being actual animals, which is a relation between the categories themselves. Indeed, much of the work in distributional semantics centers on category-denoting words, such as nouns and adjectives, and on modeling aspects of the category concepts they express. Examples include lexical entailment (Levy, Remus, et al., 2015) and other lexical relations (Baroni & Lenci, 2011) or attributes (Kelly, Devereux, & Korhonen, 2012), all of which apply at the concept level rather than at the word level (G. Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). In these works, as in ours, the reliance on language to model something extra-linguistic is typically not a fundamental one but an expedience, enabled by the availability of linguistic data and its relative ease of processing, compared to the real world.

Much less work in distributional semantics has been devoted to distributional representations of names and entities. Some exceptions are Herbelot (2015), which analyses the properties of distributional representations of person names extracted from novels; work on extracting entity attributes (Gupta et al. 2015; Guu, Miller, and Liang 2015; Hutchinson and Louwerse 2018; Louwerse and Zwaan 2009); building entity representations from annotations and knowledge bases (Bianchi and Palmonari 2017); and the relation between entities and the categories that they instantiate (i.e., instantiation, Boleda, Gupta, and Padó 2017).

In contrast, the importance of entities for the representation of category concepts has long been acknowledged in Cognitive Science. For instance, Piaget considers abstraction over members of a category the core mechanism for concept formation; prototype theory (Rosch, 1975) holds that we mentally represent categories in terms of an abstracted, prototypical member; and exemplar models (Nosofsky, 1986) hold that at least some core aspects of our concepts can be explained in terms of category members alone, the exemplars, with no further abstraction required. Our work connects with this body of research by presenting a model of category concepts that is a basic implementation of prototype theory and closely related to exemplar models (J. D. Smith, 2014).

3 Approach

3.1 Noun-based and Name-based Representations

More formally, we define two concept representations. Let $\vec{w}$ be the distributional representation we obtain for word w from some source. Then:

- **Noun-based representation** $\textsc{NounBased}(C)$ of the concept of a category C is the distributional semantic word vector of a common noun n_C that is typically used to denote category C:

$$\textsc{NounBased}(C) = \vec{n}_C$$

- **Name-based representation** $\textsc{NameBased}_{(E,F)}(C)$ of the concept of a category C is an *aggregation function* F applied to the distributional vectors of names of a set of entities E belonging to the category C:

$$\textsc{NameBased}_{(E,F)}(C) = F(\bigcup_{e \in E} \{\vec{e}\})$$

Constructing a $\textsc{NameBased}$ representation for the concept of, say, the category scientist, involves constructing distributional representations of names of actual scientists such as Emmy Noether, Albert Einstein and Marie Skłodowska Curie, and aggregating them with the function F. This method has, to our knowledge, not been considered before as category representation.

The definitions of both $\textsc{NounBased}$ and $\textsc{NameBased}$ representations are generic, which allows for many different specific models. In the present paper, we consider primarily the use of *averaging* as aggregating function. Not only is the use averaging for aggregating fairly common in distributional semantics (Bojanowski, Grave, Joulin, & Mikolov, 2017; J. Mitchell & Lapata, 2010; Westbury & Hollis, 2019), but it can also be interpreted in terms of prototype theory (Rosch, 1975): We average the distributional behavior of entities instantiating an (extra-linguistic) category to obtain a representation (prototype) of the category.

This is clearly not the only option for defining name-based category representations. The abovementioned exemplar theory (Nosofsky, 1986) provides a plausible alternative where no abstraction over stimuli takes place, and comparing concepts involves comparisons among individual exemplars (with subsequent aggregation). In this paper, we focus on the NounBased/NameBased (prototype) comparison. (We carried out additional experiments with an exemplar variant of our NameBased model along the lines just described but found essentially parallel behavior to the prototype model, with slightly lower performance across the board.)

3.2 Dataset

3.2.1 The Original Instantiation Dataset

Our NounBased and NameBased representations and the experiments we describe in Section 4 require an inventory of entities and their categories, with corresponding natural language names and nouns. For this purpose, we adapt a pre-existing dataset on the semantic relation of *instantiation*, (Boleda et al., 2017), comprising 577 categories and 4750 entities.

The structure of the dataset is shown in Table 1. Each *positive* datapoint (e, c) pairs an entity with a category it instantiates. Each positive datapoint is transformed into four *negative* datapoints as follows:

Table 1: Examples of positive entity–category pairs and four confounder types.

Type	Example 1	Example 2
Positive	George Washington – president of the US	Mumbai – city
Inverse	president of the US – George Washington	city – Mumbai
Ent2Ent	George Washington – Peter Behrens	Mumbai – Vicksburg
NotMemb-global	George Washington – river	Mumbai – statesman
NotMemb-inDomain	George Washington – astronomer	Mumbai – residential area

- Inverse: Swap the positions of entity and category, yielding (c, e).
- Ent2Ent: Replace the correct category by a different random entity e' of the same ontological domain, yielding (e, e').
- NotMemb-global: Replace the correct category c by a random wrong category c', i.e., of which e is not a member, from the global distribution of categories, yielding (e, c').
- NotMemb-inDomain: Replace the correct category c by a wrong category c'', this time sampling from the same ontological domain, yielding (e, c'').

The Inverse confounders test that the models correctly capture the asymmetric nature of the membership relation. Ent2Ent checks that the models can distinguish categories from entities and are not fooled by similarity, since entities in the same ontological domain tend to be more similar to each other. Finally, NotMemb-global and NotMemb-inDomain aim at testing that models actually learn the relation between a specific entity and a specific category, as opposed to learning to classify entities vs. categories in general (Levy, Remus, et al., 2015).

This dataset was created by Boleda et al. 2017 by extracting the positive items of the dataset from the linguistic resource WordNet (Fellbaum, 1998), where entities are linked to their corresponding categories by the instance hyponym relation (Alfonseca & Manandhar, 2002), and filtered against an existing distributional model, the Freebase model (Mikolov, Yih, and Zweig 2013, available from `https://code.google.com/archive/p/word2vec/` as `freebase-vectors-skipgram1000.bin.gz`). In cases with multiple naming variants in a WordNet synset, the FreeBase identifier was chosen that matched the longest element of the synset, assuming that it would be the least ambiguous; e.g., while Washington can refer to different entities (a person, a state, or a city), George Washington almost always refers to the former president of the USA.

The Freebase model constitutes, to our knowledge, the largest existing source for distributional representations of entity names. It contains about 1.4 million word vectors for both names (the vast majority) and category-denoting nouns. The names and nouns were drawn from the knowledge base FreeBase (Bollacker, Evans, Paritosh, Sturge, & Taylor, 2008), the largest freely available repository of entities at the time. It was trained on a 100-billion-words Google News dataset using the Skip-gram variant of the Word2Vec algorithm mentioned in Section 2, and provides 1000-dimensional representations.

The Instantiation dataset is closely related to the more applied tasks of Named Entity Recognition and Classification (NERC), in which distributional representations have found a steady place (Gouws & Søgaard, 2015; Moreno, Romá-Ferri, & Moreda, 2017; Xiao & Guo, 2014). NERC deals with the identification and classification of named entities in running text (i.e., in context) while Instantation works at the type level. Traditionally, NERC uses few categories (e.g., location, person, organization, other), but fine-grained approaches (e.g., Abhishek and Awekar 2017; Ling and Weld 2012; Shimaoka, Stenetorp, Inui, and Riedel 2017) have a granularity similar to the Instantiation dataset.

3.2.2 Our Adaptations to the Dataset

In the present paper, we use only the categories in the dataset that have at least five members. This threshold was chosen prior to our comparative experiments, motivated by the intuition that a minimal number of entity names in E is needed for the name-based representation $\textsc{NameBased}_E(C)$ of category C to be sufficiently representative of the category. We will show in Section 4.1 that a small amount of names results in good performance. When we use name-based representations computed by averaging all corresponding entities in (the training portion of) the dataset, we will denote this simply by $\textsc{NameBased}(C)$, omitting E and F.

Furthermore, we use the ontological domains extracted by Boleda et al. (2017) from WordNet's 'lexicographer file' labels (Curran, 2005; Rigau, Atserias, & Agirre, 1997) to define the fourth type of confounder above, NotMemb-inDomain, which was not present in the dataset by Boleda et al. Arguably, NotMemb-global, which draws a confounder's category at random, is not a very challenging condition (e.g., it pairs *George Washington* with *river*). In NotMemb-inDomain, confounder categories stem from the same ontological domain as the correct category, and thus is semantically more similar to the correct category (e.g., *George Washington* is paired with *astronomer*).

The resulting dataset has 159 categories and 4,180 entities (see Table 2). These combine into 4,790 datapoints, since some entities belonging to more than one category. Most of the datapoints belong to ontological domains Person and Location. The Person domain consists of popular and well known, fictional and non-fictional, historical as well as modern day people; Location contains geopolitical entities, such as countries or cities; Object mostly consists of geographical and natural entities; Communication includes literary texts but also computer programs and operating systems; Artifact covers man-made entities (including buildings, organizations); finally, Act consists of famous events. The domain Other collapses any remaining domains that had fewer than 50 categories prior to our own filtering.

These domains are also relevant for post-hoc analysis: In Experiment 1 we will contrast within-domain and between-domain comparisons of categories. We also conducted further per-domain analyses,

Table 2: Composition of the subset we use of the pre-existing Instantiation dataset.

Domain	# pairs	# entities	# categories	Example
Person	2408	2076	98	Emmy Noether, mathematician
Location	1665	1436	26	Oaxaca, city
Object	547	546	18	Nile, river
Communication	48	48	5	Hail Mary, prayer
Artifact	45	45	3	Cornell University, university
Act	43	43	4	Alamo, siege
Other	34	34	5	Paleocene, epoch
Total unique	4790	4180	159	

to see for instance whether the categories of certain domains are consistently harder or more affected by certain factors than others, but no consistent picture emerged, and in the interest of space we will not present these analyses.

Finally, we note that a few of the categories (19/159, or 12%) are in fact denoted by multi-word expressions, such as "president of the U.S.A."; the difference between these and simple nouns does not matter for our results, and we will simply include both under the header "nouns".

4 Experiments

We conducted two experiments that probe different properties of category representations. Experiment 1 assesses their ability to model human relatedness judgments betwen categories. Experiment 2 tests their ability to predict category membership, i.e., whether an entity is a member of a category (e.g., Italy – country).

4.1 Experiment 1: Category Relatedness

In Experiment 1, we sample pairs of categories from the dataset described in Section 3.2, collect human judgments of category relatedness for these pairs through crowdsourcing, and analyze their correlation with the vector similarities predicted by the two models.

4.1.1 Elicitation of Human Data

We elicit relatedness judgments for around 1000 pairs of categories, aiming for good coverage of the different ontological domains, a balance between within-domain and between-domain pairs, and reasonable coverage of the full relatedness range, from completely unrelated to closely related.

To get good coverage of the domains, we select 50 pairs of categories from each of the sparser domains (Artifact, Act, Other, Communication) and 300 pairs from each of the more populated domains (Object, Location, Person). To balance within- and between-domain pairs, we include all within-domain pairs of the sparser domains (3, 6, 10 and 10, respectively) and top up to 50 with between-domain pairs, while the more populated domains enable an even split of 150 within-domain and 150 between-domain pairs. For simplicity, we sample the between-domain pairs independently for each domain.

To get good coverage of the relatedness range, we follow previous work (Bruni, Tran, & Baroni, 2014) that samples pairs based on distributional semantic similarity, from most to least similar, divided into three bins. We performed the sampling separately within each partition. For instance, for the within-domain partition of the Person domain, requiring 150 pairs, we ranked all within-domain pairs by similarity and took the 50 most similar pairs, then another 50 from the next 100 (=50*2) pairs, and the

Table 3: Number of within/between-domain pairs of categories for which we gather human judgments.

Domain	within	between
Person	150	357
Location	150	228
Object	149	184
Communication	9	81
Act	5	60
Artifact	2	56
Other	9	48
total	474	507

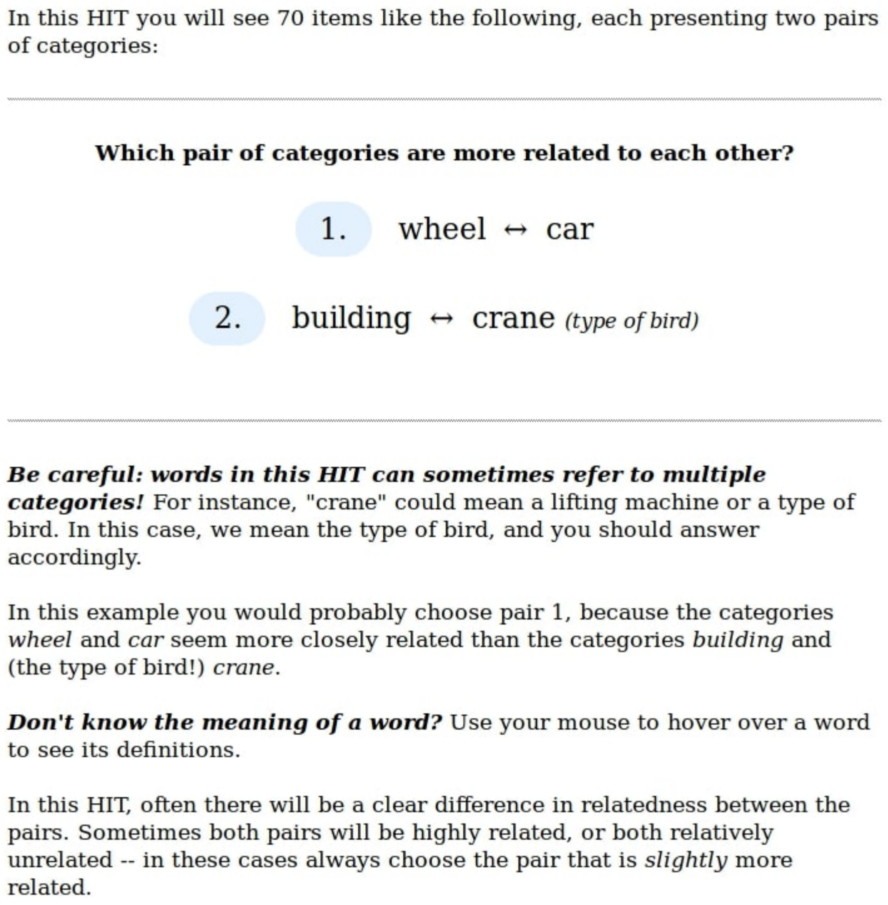
last 50 by randomly sampling from all remaining (mostly dissimilar) pairs. For the smaller partitions, we either included all available pairs, or sampled using only two bins instead of three (i.e., most similar pairs and the rest). After sampling, we remove duplicate pairs (which arise from independently sampling the between-domain partitions), resulting in 981 unique pairs of categories for which we elicit human judgments of category relatedness. Table 3 summarizes the number of pairs available for each partition. Note that the numbers in the right-hand column do not add to the total because each between-domain pair contributes to the counts in two rows.

Figure 1 shows the set-up of our experiment, with the instructions given to participants. Again following Bruni et al. (2014), we present two pairs at a time and ask which one is more related, and aggregating the resulting binary judgments to obtain the proportion of times the target pair was judged to be more related than another. This results in a score between 0 and 1 for each of our original 981 category pairs. Since comparing all pairs of pairs is infeasible, we pair each category pair with 50 randomly picked pairs, resulting in a total of 49,050 binary judgment tasks. Below we will evaluate our models by comparing their cosine similarities to the resulting, aggregated human relatedness scores. (We also tested the models directly against the raw, comparative judgments, resulting in the exact same significant patterns. For the sake of conservativeness, we will use only the method of Bruni et al. 2014 below.)

We obtained the judgments through crowdsourcing, setting up our task using the Ibex framework ('internet-based experiments', `https://github.com/addrummond/ibex`). We recruited participants on Amazon Mechanical Turk, sending them to our experiment hosted on IbexFarm (`http://spellout.net/ibexfarm`). Our data collection design was approved by the Ethical Committee of Pompeu Fabra university. Participants were given instructions followed by 70 items to judge (15 of which were control items, see below). Both the order of pairs in an item and the order of items in a task were randomized for each participant. A single participant could do at most 6 of these 55-item tasks, thus covering at most one third of the category pairs.

We composed 90 quality control items (15 for each of the six tasks a single participant may work on) by pairing 90 pairs of categories that were close together in distributional space with 90 pairs that were far apart. We manually went through the resulting control items to verify that the intended pair was indeed clearly more related and further removed 9 fillers post-hoc (those where less than 90% of participants agreed with our own judgment). We discarded 86 submitted tasks where the accuracy on the remaining fillers was less than 85%. We ran another crowdsourcing round in order to mostly fill the resulting gaps, from which we discarded again 7 due to the 85% threshold. Ultimately, we obtained at least 48 judgments per category pair, 50 for most.

An important methodological consideration is that we aim to obtain category relatedness, instead of word relatedness, in contrast to previous work (Agirre et al., 2009; Hill, Reichart, & Korhonen, 2015;



In this HIT you will see 70 items like the following, each presenting two pairs of categories:

Which pair of categories are more related to each other?

1. wheel ↔ car

2. building ↔ crane *(type of bird)*

Be careful: words in this HIT can sometimes refer to multiple categories! For instance, "crane" could mean a lifting machine or a type of bird. In this case, we mean the type of bird, and you should answer accordingly.

In this example you would probably choose pair 1, because the categories *wheel* and *car* seem more closely related than the categories *building* and (the type of bird!) *crane*.

Don't know the meaning of a word? Use your mouse to hover over a word to see its definitions.

In this HIT, often there will be a clear difference in relatedness between the pairs. Sometimes both pairs will be highly related, or both relatively unrelated -- in these cases always choose the pair that is *slightly* more related.

Figure 1: The instructions shown to participants when entering our crowdsource task.

Rubenstein & Goodenough, 1965). Since the same noun can typically denote multiple distinct categories, we need to indicate to participants exactly which category we want them to judge. In fact, participants tasked with judging word relatedness may ultimately be using category relatedness to do so (Bach, 2002), and a frequent observation in studies on word relatedness is the 'good subject effect' (Nichols & Maner, 2008), as participants are typically willing to adapt their interpretation of words to context: "bank" can be found to be similar both to "river" and to "money", by interpreting it as a different category each time. To disambiguate, we add phrases in parentheses, informed by WordNet, to all words we consider potentially ambiguous (such as "type of bird" in Figure 1). This resulted in adding disambiguating phrases to 29 categories (18%), for instance muse (mythology), plateau (geography) and star (astronomy). Participants were further encouraged to hover over any word to view its definition, which we again extracted from WordNet.

4.1.2 Results of Experiment 1

We first consider the reliability of the collected dataset. Assessing inter-annotator agreement is tricky, since annotators saw different items (except the fillers). To obtain an estimate, we collected additional parallel judgments from 8 annotators for a subset of 55 pairs plus 15 fillers. As above, we removed annotators who scored less than 85% correct on the fillers (three). For the remaining 5 annotators, on the 55 non-filler items, Krippendorff's Alpha (Krippendorff 1980; we use the Python `krippendorff` package) is 0.41 (vs. 0.94 on the fillers). This is only moderate, but we must take the nature of the task into account. Manual inspection of the 18 of 55 cases where the annotators were divided revealed these to consist, without exception, of two roughly equally related (e.g., bishop–epistle/dictator–war) or

Table 4: Main results of Experiment 1: Spearman correlation coefficients, significant differences bolded.

	all	within-domain	between-domain
number of pairs	981	474	507
NounBased	0.56	0.57	0.64
NameBased	**0.74**	**0.67**	0.69

unrelated pairs (e.g., thoroughbred–muse/plateau–apostle), making the forced binary choice arbitrary or extremely subjective. Since pairs appear in much clearer comparisons, and there are 50 comparisons for each pair, the effect of this kind of noise should disappear, and we conclude that inter-annotator agreement is sufficient, given the nature of the task, to warrant analysis of the dataset.

To assess the performance of the models, we compute the correlations between the cosine similarities according to our two models and the aggregated human relatedness scores. Table 4 clearly shows that the NameBased model provides a better fit to human judgments of category relatedness. Since the two kinds of score have different distributions, we use Spearman's correlation coefficient. Scatter plots are available in Appendix A.1 (Figure 3). The leftmost column reports correlations computed on all pairs of categories, showing that the NameBased model obtains a strong correlation with human judgments ($\rho = 0.74, p = 1\mathrm{e}{-170}$) and beats the NounBased model ($\rho = 0.56, p = 3\mathrm{e}{-82}$) by quite a margin. The correlations are significantly different ($t = 8.3$, $p = 2\mathrm{e}{-16}$) using a two-tailed dependent correlations test from Steiger 1980 (described in Howell 2012, p.287; we used `https://github.com/psinger/CorrelationStats`, which is based on the R package paired.r).

The two right-hand columns of Table 4 show that the advantage of the NameBased model over the NounBased model is in pairs of categories from the same ontological domain (within-domain; $t = 3.1$, $p = 0.002$), with instead no significant difference in different domains (between-domain; $t = 1.5$, $p = 0.13$). Since differences between categories from the same domain are likely to be more subtle than those between categories of different domains, this suggests that the NounBased model is more coarse-grained than the NameBased model. We take the fact that the difference between the NounBased and NameBased models overall (leftmost column) is greater than in either the within-domain or between-domain subset to mean that the NameBased model is also better at ranking between-domain and within-domain pairs relative to each other.

The quality of the NameBased model is expected to depend on the particular selection of entities used, i.e., the set E in the definition of NameBased$_{(E,F)}$ in Section 3. To investigate this, Figure 2 shows the performance (Spearman correlation with human judgments) of the NameBased$_{(E,F)}$ model as the number of entities in E goes from 1 to 5, with a final jump to the maximum number of names available. The horizontal dashed line indicates the correlation of the NounBased model (0.56; see Table 4). As can be seen, three names per category suffice for the NameBased model to surpass the NounBased model, and four names to surpass it even on the unambiguous categories, where the NounBased model has less of a disadvantage.

The vertical lines in Figure 2 show the standard deviation in correlation coefficients, obtained by sampling for each category, for every size from 1 to 5, up to 50 different name sets of that size (limited by availability of entities). The standard deviations are quite small; therefore, we conclude that the NameBased model's overall performance (i.e., averaged over all categories) does not depend much on the particular entities chosen; the number of entities is far more important.

Nevertheless, for particular categories entity representativeness can play a large role. For instance, the pair of categories whose relatedness is most overestimated by the NameBased model is (surgeon, siege). The surgeons in the dataset we use are Alexis Carrel, Walter Reed, William Beaumont, William Cowper, James Parkinson, and Joseph Lister. It turns out that Alexis Carrel was involved in World War

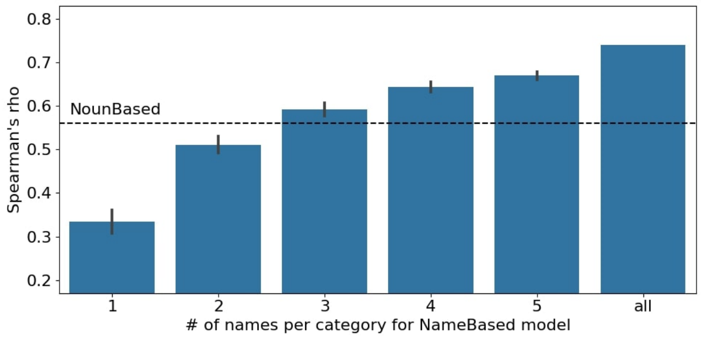


Figure 2: Correlation coefficients of the NameBased model based on different numbers of names, compared to the NounBased model (dashed line). Vertical lines show standard deviations.

I and Walter Reed and William Beaumont were members of the US Army Medical Corps, which means that at least half of the surgeons had a military connection, a category closely related to siege. This may reveal a more general shortcoming of relying on public knowledge bases such as WordNet for entity names: named entities who are famous enough to be included in such a resource tend to be historically influential figures, which may skew the entity's contexts of occurrence. To clarify: the challenge here is not that the same entity is typically a member of multiple categories, or that membership in one category can correlate with membership in another; the challenge is, rather, that the correlations between categories in the knowledge base should be representative of correlations between categories in the real world.

Summing up, on predicting aggregated human judgments of category relatedness, the NameBased model is superior to the NounBased model, as long as sufficiently many entities are used (upwards of 3 overall) and provided that the entities do not happen to share irrelevant properties (such as the surgeons' relations to the military). As mentioned in Section 3, we also tested a version of the name-based model based on exemplar theory (Nosofsky, 1986), which, instead of computing the cosine similarity between the entity averages (or prototypes) of two categories, computes the average of all cosine similarities between pairs of entities from the two categories. With only a single entity per category, the prototype and exemplar models are equivalent; with multiple entities per category the exemplar model performs consistently slightly worse than the NameBased prototype model, but with the same qualitative pattern (performance significantly above NounBased for $n \geq 4$). This is in line with studies which indicate that for current representation learning approaches, the non-linear decision boundaries of exemplar models may not provide advantages over prototype models (Sikos & Padó, 2019).

4.2 Experiment 2: Category Membership

This experiment tests which representations, NounBased or NameBased, are better for predicting *category membership*: whether an entity is a member of a category (e.g., Italy – country). We do so by training neural network classifiers to predict category membership using either NounBased or NameBased representations of categories as inputs, the reasoning being that better representations of category concepts will lead to better classification accuracy. Training a classifier with certain vector representations as input is a standard method to test what information those vector representations contain (see Sommerauer and Fokkens 2018 for a recent linguistic example; see Günther et al. 2019 for discussion).[1] We largely follow the dataset and methodology of Boleda et al. (2017), who first framed category membership prediction as a Machine Learning task, modulo the adaptations described

[1] Thus, the classifiers we train in this section are not to be considered part of the NounBased or NameBased models, but, like the correlation analysis in Experiment 1, a method for diagnosing what information the two kinds of representations contain.

in Section 3.2.

We combine positive pairs with different types of confounders (cf. Table 1) to obtain different datasets which enable a systematic investigation of different aspects of the task, in particular teasing apart the general distinction between instances and categories and the membership relation proper. We build one dataset for each confounder type (total: four), as well as two Union sets that combine different types of confounders. Both unions include Inverse and Ent2Ent confounders; Union-global adds NotMemb-global, and Union-inDomain adds NotMemb-inDomain. These Union datasets consists of 19160 data points each, of which 25% positive examples. The Union datasets are more challenging than the other datasets that they require models to distinguish positive examples from confounders of different types. Therefore, we will focus on these two datasets here, and report results on the remaining datasets in Appendix A.2.

4.2.1 Training and Testing Regime

Like Boleda et al. (2017) for the original dataset, we randomly split our modified version into training, validation and test sets (80%, 10%, and 10% respectively). Maintaining this division for fitting and evaluating models is, however, not enough to be able to test whether a model can truly generalize: the related task of hypernymy detection has been shown to suffer from the problem of *memorization* (Levy, Remus, et al., 2015; Roller et al., 2014) — a problem not considered by Boleda et al. (2017).

Memorization was first found in hypernymy, in which models were found to learn that certain words (such as *animal*) make good hypernyms because they are general, irrespective of the word they are paired with. This can lead to overoptimistic evaluation results even if train and test *pairs* are disjoint, as long as *words* are reused in different pairs in the two partitions. Thus, to obtain meaningful evaluation results, the train and test vocabularies must be kept completely disjoint. To this end, we adopt the methodology of Roller et al. (2014), which ensures zero lexical overlap between training, validation, and test sets. Specifically, we split the test set into many equal-sized test folds and remove overlap of each fold with the training and validation data: For example, if (Mumbai, city) occurs in the current test fold, then all pairs containing either Mumbai or city are removed from the corresponding training and validation data. We choose the number of test folds so that the average size of the training set, after removing the lexical overlap, is still 90% of the original training data (fewer, and therefore larger, test folds would lead to excluding more training data; smaller test folds, e.g., leave-one-out (Roller & Erk, 2016), would increase computational load substantially). This results in 83 test folds.

4.2.2 Models

We again compare the NameBased and NounBased representations as defined in Section 3, using Freebase word vectors as representations of entity concepts and category concepts. The NounBased and NameBased representations are fed as input to a classifier. This time, $\text{NameBased}_{(E,F)}(C)$ for the category C is constructed with E as the set of all entities of the category C in the training set. This means that no entity that occurs in the test set is ever used in the construction of the NameBased representations.

For this experiment, we make sure that all vectors are L_2-normalized (i.e., scaled to length 1), and in addition L_2-normalize the NameBased representations, so that the classifier cannot recognize NameBased representations solely by virtue of their length. (L_2-normalization was not necessary in Experiment 1 because we used cosine similarities, which reflect vector directions but not length.) Furthermore, as is common to facilitate training, we scale the input values column-wise so the values for each input dimension fill the [-1, 1] range.

Table 5: Main results on the category membership task, highest scores in bold.

				NounBased			NameBased		
Measure	Dataset	BL_{Freq}	BL_{Pos}	Cos	NN-1HL	NN-2HL	Cos	NN-1HL	NN-2HL
Micro F_1	Union-global	0.25	0.40	0.43	0.74	0.77	0.59	**0.85**	0.75
	Union-inDomain	0.25	0.40	0.41	0.51	0.65	0.55	**0.76**	0.68
Macro F_1	Union-global	0.20	0.43	0.41	0.38	0.45	0.45	**0.65**	0.53
	Union-inDomain	0.21	0.43	0.40	0.21	0.37	0.42	**0.59**	0.47

For classification, we use a feed-forward neural network (NN) (we employ the Keras toolkit, `https://keras.io`). We report results on neural networks with one hidden layer (NN-1HL) and two hidden layers (NN-2HL). Inputs are pairs $(\vec{v}, \vec{w})$, where v is an entity and w a category in the case of positive data points, but either could be an entity or category depending on the type of confounder. Following the best-performing setup of Boleda et al. (2017), we represent each pair (v, w) by a concatenation of their vector representations $\langle v_1, \ldots, v_n, w_1, \ldots, w_n \rangle$, where for categories we use either the NounBased or NameBased representation.

The model is trained to predict whether the first element is a member of the second element (which also implies that the first element is an entity, and the second a category). The models are trained with mean cross-entropy as loss function, using Adadelta optimization (Zeiler, 2012) to a maximum of 2000 epochs with early stopping (we found that models typically converge at 50–100 epochs). All hidden layers use `tanh` as activation function; the output (classification) layer uses `softmax` as activation function. The number of units in each hidden layer of the NN models is optimized for each model separately through hyperparameter search on the values 5, 10, and from 50 to 800 with step size 50. To reduce overfitting, we introduce a dropout layer in front of each hidden layer with a standard dropout value of 0.5 (Baldi & Sandowski, 2013). We performed additional experiments (3–4 hidden layers, additional regularization) but did not find any benefits.

4.2.3 Evaluation and Baselines

Our primary evaluation measure is micro F_1-score on the test partition for comparability to Boleda et al. (2017). Since this measure focuses mainly on larger categories, and our categories vary greatly in size (the Union-inDomain test set has between 1 and 152 data points per category: mean 10.3, SD 18.9), we also report macro F_1-score: We compute F_1 for each category separately, then average these scores.

We compare the neural network models to two baselines and a simple similarity-based model. The frequency baseline BL_{Freq} assigns the positive class randomly with the true class probability. The positive baseline BL_{Pos} always assigns the positive class. The similarity model, Cos, computed for NounBased and NameBased representations separately, establishes what semantic similarity on its own can achieve: It classifies each potential entity-category pair based on cosine similarity between the representations of the two members of a pair, where the classification threshold is chosen to maximize micro F_1-score on the training and development set combined. Since cosine is known to reflect relatedness and to not distinguish between different kinds of semantic relations (e.g., hypernymy vs. synonymy; Baroni and Lenci 2011), we see this model as an informed baseline (see Appendix A.3 for analysis of the cosine baseline).

4.2.4 Results of Experiment 2

Table 5 shows the main results on the category membership task. In both subsets and according to both evaluation measures, models that use name-based representations beat those that use noun-based representations by a large margin ($p < 0.001$, determined by bootstrap resampling). The best NameBased model, NN-1HL, obtains between 8 and 25 F_1-score points more than the best NounBased model, NN-2HL. It also has reasonable to good absolute performance, with 0.76–0.85 micro F_1-score, 0.59–0.65 macro F_1-score. (The lower results for macro F_1-score are expected, since smaller categories are generally harder to model.) While the neural network models beat the baselines across the board, the cosine models Cos only clearly beat them for the NameBased model in micro F_1-score evaluation. We interpret this to say that (a) the name-based category representations are already by themselves more informative than the noun-based category representations; that (b) in an unsupservised setting, this effect is only visible for high-frequency categories; but that (c) a supervised setting can carry this effect over to lower-frequency categories. In the macro-F_1 evaluation, it is furthermore striking that one or both NounBased NN models end up worse than the Cos baseline, indicating that the more powerful NounBased NN models struggle with the less frequent categories, in a way that the corresponding NameBased models do not. The NameBased representations, therefore, seem to be easier for a neural network to generalize over.

When we compare the different neural models, we note that when we evaluate with micro F_1-score, for NameBased representations, the one-layer model (NN-1HL) is superior. In contrast, for NounBased representations, it is the more complex two-layer model (NN-2HL). This suggests that not only is there more useful information present in the NameBased representations, it is also more directly retrievable: Identifying the membership relation on the basis of NounBased representations requires a more complex transformation.

Altogether, the foregoing results show that the NameBased representations contain information that is more useful for modeling membership, an essential aspect of our knowledge of categories. This aligns with the observed superiority of NameBased representations in Experiment 1, further corroborating our hypothesis that NameBased representations are a more adequate model of category concepts than NounBased representations.

5 Analysis and Discussion

Both experiments support our hypothesis that the NameBased model is empirically a better model for category concepts. We now reflect on this finding from three angles, in three subsections: ambiguity of nouns as a possible cause for the superiority of the NameBased model, model complexity, and explanatory potential.

5.1 The Impact of Ambiguity on the Noun-based Model

We started this paper by arguing that a factor that contributes to the attractiveness of names as distributional anchors is that they are comparatively ‘rigid’ (Kripke, 1980), in that they are almost universally used to refer to a particular individual, and therefore exhibit a superior alignment between words and concepts compared to common nouns. If this is true, we would expect the empirical benefit that we find for the NameBased models to correlate with the extent of ambiguity of nouns, and the following analysis looks for signs of ambiguity-related effects in Experiments 1 and 2.

Since the degree of ambiguity of a wide range of nouns is difficult to estimate consistently, we look not at ambiguity itself, but at an important consequence of it: that in some cases the noun’s predominant sense may not coincide with the target category to begin with. We expect the NounBased model to struggle in particular when there are such mismatches, because word vectors, being data-driven, prominently mirror the word’s predominant sense (Arora, Li, Liang, Ma, & Risteski, 2018; McCarthy, Koeling, Weeds, & Carroll, 2004). Recall that in the crowdsourcing-based construction of our dataset, we

Table 6: Analysis of Experiment 1 in terms of ambiguity: Spearman correlation coefficients between human judgments of relatedness and model predictions.

	all	matched	ambiguous
number of pairs	981	626	355
NounBased	0.56	0.62	0.45
NameBased	0.74	0.73	0.75

provided a disambiguating phrase for a noun whenever we felt that it was in doubt that the participants' interpretation matched out target concept (e.g., "star (astronomy)"). Erring on the side of caution, this applied to 29 categories (18%). We reuse this distinction for the present analysis as follows. Note that, by following the WordNet senses, we do not distinguish between polysemy (related senses) and homonymy (unrelated senses). This makes our analysis below conservative as far as homonymy goes: if we find an effect of non-predominent senses due to polysemy/homonymy, we certainly expect an effect for homonymy alone (we leave testing this expectation to future work).

To analyze the results of Experiment 1, we label those cases as 'matched' where both nouns of a pair have predominant senses that match our target concepts. We label the other cases, where we provided a disambiguating phrase for at least one noun, as 'ambiguous'. Table 6 shows the correlation coefficients of Experiment 1 for the ambiguous and unambiguous subsets (the first column, with overall results, is identical to Table 4). This reveals a big performance gap for the NounBased model between the two subsets (0.62 for matched and 0.45 for ambiguous cases, respectively; $p < 0.001$, determined by bootstrap resampling). By contrast, the NameBased model's performance does not change ($p \approx 0.6$), which is as expected because this model is not sensitive to properties of the nouns. The difference in correlation coefficient between matched and ambiguous is significantly greater for the NounBased model (by on average 0.18, $p < 0.001$, again by bootstrap resampling). Furthermore, note that the NameBased model is still superior to the NounBased model even in cases where the target category and the noun's predominant sense match (0.73 vs. 0.62; $t = 4.7$, $p = 2.5\text{e}{-6}$; as in Section 4.1 using the two-tailed dependent correlations test from Steiger 1980). This is expected, too: even for nouns whose predominant sense is the target category, other senses can still affect their distributional representation, as do the many other factors to which nouns are more susceptible (e.g., metaphor or connotation).

Let us look through the same lens at the results of Experiment 2, based on the 'matched'/'ambiguous' division. We focus on the toughest dataset, Union-inDomain, and the best models for each representation, namely NN-2HL for NounBased and NN-1HL for NameBased. (The trends we find there are consistent with those of the other neural network models.) Table 7 shows that we find a decrease in performance when the predominant sense of the noun does not match the intended category, as in Experiment 1. Here, both per-data-point and per-category scores support the same conclusion. As in Experiment 1, we find that the NounBased model struggles with the ambiguous cases: 0.42 vs. 0.67 in micro F_1-score ($p < 0.001$, determined with bootstrap resampling), and 0.29 vs. 0.39 in macro F_1-score ($p \approx 0.045$). An initially puzzling difference with Experiment 1 is that, in Experiment 2, we also find a slightly worse performance on ambiguous cases for the NameBased model: 0.68 vs. 0.77 in micro F_1-score, 0.54 vs. 0.60 in macro F_1-score. Although these differences for the NameBased model are themselves not significant ($p \approx 0.2$ and $p \approx 0.075$, respectively); they are substantial enough for there to be, unlike in Experiment 1, no significant difference in the degrees to which the NounBased and NameBased model are affected by matched vs. ambiguous ($p \approx 0.081$, again by bootstrap resampling; cf. Gelman and Stern 2006). Since the NameBased model by definition cannot be directly sensitive to linguistic properties of the noun, this tendency (if real) must have a different explanation. Indeed, we know from Experiment 1 that the number of entities matters for the quality of NameBased repre-

Table 7: Analysis of Experiment 2 in terms of ambiguity: F_1-scores of the best models on the UnionInDomain test set. Highest scores in bold.

		all*	matched	ambiguous
	number of data points:	1437	1316	121
	number of unique categories:	140	114	26
	proportion data points positive:	0.33	.34	.31
Micro F_1	NounBased	0.65	0.67	0.42
	NameBased	**0.76**	**0.77**	**0.68**
Macro F_1	NounBased	0.37	0.39	0.29
	NameBased	**0.59**	**0.60**	**0.54**

*These are only all data points that involve a category (i.e., excluding Ent2Ent), because we are interested here in the effect of ambiguity of category-denoting nouns.

sentations, and ‘ambiguous’ categories tend to have much fewer entities: 13 entities on average (std. 9) compared to 34 entities on average for the ‘matched’ nouns (std. 69). We take this as at least a partial explanation for the performance dip of the NameBased model for the ‘ambiguous’ subset.

From the analysis of both experiments combined, we conclude that the inclusion of nouns that may not predominantly denote the target category gives the NameBased model part of its advantage over the NounBased model. Although this may seem like an ‘unfair’ advantage that one could try to blame instead on the choice of nouns in our dataset, we think that would merely reframe the fact that nouns vary in how clearly they denote a given category, which is one aspect of the complex mapping between language and the world to which names are simply less susceptible.

5.2 Model Complexity

We now move to a different level of consideration, namely the inherent complexity of the NameBased vs. the NounBased model. It may seem that the NameBased model has more information about categories built into it than the NounBased model, namely, information about which entities are members of which categories. If true, this would mean that the superior performance of the NameBased model, compared to the NounBased model, could in principle be a largely uninteresting consequence of the model’s being endowed with more information about the task at hand (rather than different information). (To clarify, the issue here is not circularity: Experiment 2 tested each model’s ability to categorize new, unseen entities.) There is also a more practical counterpart of the same concern: NounBased representations are easy to come by (category nouns are covered by many distributional models), whereas NameBased representations require some work (finding entities, averaging their names’ vectors; we will address the related concern that not every category has named entities to begin with further below, in Section 6, because this is a matter not of model complexity but generalizability).

With regard to this question, it is important to acknowledge that *both* models presuppose a certain mapping between language and the world: Both models rely on linguistic information, namely distributional representations, in order to model what are essentially extra-linguistic judgments, i.e., judgments about the structure of the external world (and this is a common expediency, see Section 2). More precisely, the NounBased model presupposes a direct mapping between nouns and categories, while the NameBased model presupposes a more indirect mapping, from names to entities and from entities to categories. Now, while it may indeed seem easier to establish the direct mapping (for the NounBased model) than the indirect mapping (for the NameBased model), we argue that it only seems this way, and this is in fact an example of overlooking implicit reliance of representations in cognitive modelling on prior human judgments (Jones et al., 2015; Westbury, 2016). The reason is that we happen to have intuitive knowledge of which English nouns typically express which categories – and in cases of doubt we can reference a dictionary. But for a fair model comparison, this knowledge must be considered a prerequisite of the NounBased model, just as information about names and entities is a prerequisite for the NameBased model. More generally, the practical ease or difficulty of setting up a model is not a reliable indicator of the relative complexity of that model from a more theoretically informed, cognitive perspective.

Still, the NameBased model may seem to require more effort than the NounBased model, in that it requires multiple names per category (recall that models based on one or two names performed poorly in Experiment 1), whereas the NounBased model needs only a single noun per category. However, it is not clear whether the effort required to be able to choose the right noun for a category is bigger or smaller than the effort required to be able to link multiple names to entities of the category. The latter requires only the ability to name examples of a category, which does not require knowledge about the category other than some examples of its instantiation relation; the former requires more complete knowledge (e.g., a definition) of the category, or at least enough to determine whether a given noun is sufficiently reliably used to express exactly that category. And since nouns are generally both more frequent and versatile than names, the latter can be quite a challenge. To illustrate: if you say "Mao Tse Tung" and "Jiang Zemin", then you may be able to add "Hu Jintao" even without knowing the specific title of their category ("chairman of the Communist Party of China"), or what being an instance of that category entails exactly.

Another, independent issue related to model complexity is the amount of information contained in the distributional representations of nouns vs. names. We estimated the frequencies of the nouns and names of our dataset and found category-denoting nouns to be on average around 500 times more frequent than all names of entities of that category combined.[2] Thus, nouns arguably contain more distributional information than names. At the same time, just three names per category were enough to beat the NounBased model. Thus, in relative terms, the names are substantially more informative for category representation than the nouns. This also relates to the possible impression that the NameBased model is more complex in that it requires averaging over multiple instances: since computing distributional representations itself involves averaging over occurrences, arguably a lot more averaging went into the NounBased model than the NameBased model – it is just more explicit in the latter.

Summing up, the appearance that the NameBased model has more built-in information is arguably false (and we presented some arguments for the inverse claim, i.e., that it is the NounBased model that has more information built-in). We therefore conclude that the NameBased model's empirical superiority in both experiments is not due to it having more information about categories built-in than the NounBased model, but due to it being based on different information. The experiments therefore convincingly show that the aggregated distributions of names are genuinely more representative of category concepts than the distributions of nouns.

5.3 Explanatory Potential as Models of Concept Acquisition

As mentioned in Section 2, distributional semantics has been argued to be not just a useful tool, but an explanatory, cognitively plausible model of lexical acquisition and lexical meaning: distributional representations are the result of applying general purpose learning mechanisms to words across contexts, similarly to at least some of the mechanisms by which humans learn the meanings of words (e.g., Günther et al. 2019; Jones et al. 2015; Mandera et al. 2017). Here we address a typically overlooked question, namely whether this explanatory potential carries over to the use of distributional semantics as a model

[2] As the full 100 billion word Google News corpus on which the Freebase model was trained is not available, we computed these word frequencies on the 860 million word sample 2012/2013 from `http://www.statmt.org/wmt14/training-monolingual-news-crawl`.

not of lexical meaning but of category concepts, which is what we are aiming to model in this paper. For although language is known to facilitate category concept acquisition (e.g., Casasola, Bhagwat, and Burke 2009; LaTourrette and Waxman 2019; Xu 2002) and more abstract concepts may even be learned predominantly through linguistic means (Borghi and Binkofski 2014), most category concepts are acquired predominantly from exposure to category members in the external world.

A standard view in the literature on distributional semantics is that language can be used as a sufficiently good proxy for experience of the external world, to learn about at least some aspects of the external world. This is because there is an (imperfect) correspondence between statistical regularities in language and statistical regularities in the world (e.g., Barsalou, Santos, Simmons, and Wilson 2008; Louwerse 2008, 2018). We subscribe to this view, and thus regard both the NameBased and NounBased models as models of concept acquisition from (an approximation of) extra-linguistic experience. Note that to subscribe to this view is not to deny or bypass the symbol grounding problem (Harnad, 1990; Searle, 1980); we briefly return to this in Section 6.

Under this interpretation of the models, a crucial difference between the two is in the units of experience over which they abstract to form category concepts. For the NameBased model these units are occurrences of category members – because name tokens in text are a proxy for occurrences of the category members to which they refer. More precisely, the name-based model implements concept acquisition as abstraction over occurrences of category members: it first computes category member representations (i.e., name vectors) and then averages them to obtain a category prototype. We explicitly motivated this in terms of prototype theory (Rosch, 1975), and the more general notion of abstraction over category members aligns also with the broader literature on concept acquisition, where the vast majority of experiments involve exposing humans to several category members and then testing their category induction (e.g., Casasola et al. 2009; LaTourrette and Waxman 2019; Xu 2002).

The situation for the NounBased model is less clear. In using distributional representations of nouns to model category concepts, it must explain the acquisition of category concepts in terms of abstraction over units of experience that correspond to occurrences of the category nouns. Although noun occurrences can involve category members, chiefly in categorization acts mirrored in referential noun phrases ("The scientist entered the lab") and predicative uses ("Marie is a scientist"), nouns are, on the whole, less tightly connected to category members than names. One reason is the ambiguity of nouns, which obscures both – at the type level – the relation between nouns and categories (section 5.1) and – at the token level – the relation between specific uses of nouns and category members. An additional reason is that nouns also occur in so-called generic uses ("Scientists need grants"), which do not involve reference to a category member, but correspond to quasi-definitional category information: while a great source of category information, such noun tokens do not serve as proxies for category members. Although generic category information plays an important role in less entity-centric views of concepts, such as the classical definitional approach and the 'theory' theory, it does not feature as prominently in most contemporary work on concept acquisition (an exception being extensive work on the role of definitions in the acquisition of technical concepts, e.g., mathematics; Vinner 2002).

In sum, we conclude that the NameBased model is more straightforwardly embedded in the current cognitive literature on concept acquisition than the NounBased model, within which further questions arise about how the various aspects of category information that nouns provide are to be integrated into a category concept model. Note that neither of the two models as yet provides a complete model of category acquisition from scratch, i.e., category induction (as in Nosofsky 1986; Rosch, Simpson, and Miller 1976); both NameBased and NounBased models currently start with a fixed inventory of categories and a set of observations per category (namely instances and noun occurrences, respectively). This is but a choice of scope, not an intrinsic limitation.

6 Conclusion and Future Work

Much computational and cognitive work on distributional semantics relies on the use of language as a proxy for experience of the external world, and employs distributional representations of category-denoting nouns as a proxy for category concepts. Because names are more rigidly connected to the external world than nouns (Kripke, 1980), our primary objective was to test the hypothesis that names of category members provide a better distributional proxy for category concepts than the category-denoting nouns of existing approaches. More precisely, we used name vectors as a proxy for entities (Gupta et al., 2015; Herbelot, 2015) and averaged them to obtain a category prototype (Rosch, 1975).

Corroborating our hypothesis, we have shown that such NameBased representations, even based on as few as three names, outperform NounBased representations in two computational experiments, simulating human judgments on category relatedness and category membership, where the latter also showed its ability to generalize to new, uncategorized instances. (We also considered a variant of the NameBased model based on exemplar theory (Nosofsky, 1986), which while slightly worse than the NameBased model still beat the NounBased model by a large margin.) Further analysis in Section 5.1 confirmed that the existence of other, dominant word senses negatively impacts the NounBased model, as expected. We also reflected on the relative complexity (Section 5.2) and explanatory potential (Section 5.3) of the two models, arguing for no substantial difference in the former, and a methodological advantage for the NameBased model with regard to the latter, namely, that the NameBased model aligns more closely with the cognitive literature on concept acquisition, which emphasizes the role of exemplar occurrences in concept acquisition. The NounBased model instead provides a wider range of kinds of information about concepts (e.g., definitional information arising from generic usage) which remains important to explore but is less well understood.

Two clarifications about the scope of our argumentation are in order. First, we focus on the use of distributional semantics as a model of extra-linguistic category concepts, and our criticism of the NounBased model does not apply to its (also very common) use as a model of the lexicon. Unlike a category representation, an adequate lexical representation of a noun should reflect its various senses and connotations rather than ignore them. As such, we expect that NounBased representations show an advantage over NameBased representations when it comes to modeling the lexical knowledge about nouns (Baroni & Lenci, 2011). Interestingly, when conceived of as a model of word representations instead of category concepts, the NounBased model is itself an instance-based model: just as the NameBased model represents a category as an abstraction over (names of) category instances, the NounBased model represents a word as an abstraction over word instances.

Second, our arguments do not bear directly on the symbol grounding problem (Harnad, 1990; Searle, 1980). Although we argue that names are better proxies for the external world than nouns, they remain proxies. This is true also for multi-modal approaches that enrich text-derived word representations with visual contexts (e.g., Feng and Lapata 2010) – they are still fundamentally representations of words (Westera & Boleda, 2019). (Note that we are not talking here about entirely non-linguistic distributional approaches, such as the use of embeddings in computer vision (Simonyan & Zisserman, 2015)).

This article is the first, to our knowledge, to systematically investigate the role of entities in the representation of category concepts in distributional semantics. Some obstacles for a more general application of our proposal are that most category instances are not named, that most named instances do not appear in corpora, and that finding category instances to begin with currently depends on human-curated knowledge bases (though recall from Section 5.2 that the NounBased model relies on similar knowledge, albeit implicitly). Although our primary aim was to test a theoretically motivated hypothesis, we also view our paper as a proof of concept that entity-based representations are worth pursuing. Future work should seek to apply our proposal more widely, by exploiting recent developments in computational linguistics. We briefly review four promising avenues, although we stress that, in our view, having better category concept representations of even a subset of the categories can be useful, for instance to gain a better understanding of those categories.

First, although distributional semantics itself is extremely data-hungry, current research has developed methods that can learn more from less data, for instance by detecting and exploiting informative contexts (Herbelot & Baroni, 2017). This direction could both expand the set of named entities for which sufficiently adequate distributional representations can be computed, thus widening the coverage of the method, and improve its general cognitive plausibility.

Second, while this article has concentrated on the use of entities to characterize category concepts, future research should investigate to what extent subcategory-denoting nouns can be used for representing the super-category. For instance, the category furniture could be modelled by averaging the distributional representations of "chair", "table", etc. This could allow for the representation of categories which lack (public) named entities (e.g., individual chairs and tables typically don't have names). Although – as we have argued in this article – nouns are more problematic than names for modeling category concepts, two consideration make this particular avenue promising: (a) the aggregation of several subcategory-denoting nouns, as we did with entities to form a prototype, could smoothen out the idiosyncrasies of each noun and focus the representation on aspects related to the supercategory; and (b) the more specific subcategories can be expected to show less ambiguity (Gilhooly & Logie, 1980). Note however that an approach based on subcategory-denoting nouns would still, like the simple NounBased model, diverge from the predominant, entity-based view in the literature on concept acquisition (Section 5.3).

Third, recent neural network models provide contextualized representations of words (Devlin et al., 2019; Peters et al., 2018) which differ crucially from traditional distributional semantic representations in that the same word is assigned a different vector depending on the context in which it occurs, and, thereby, depending on, e.g., the sense in which it is used (Wiedemann, Remus, Chawla, & Biemann, 2019). Although contextualized representations are now widely used in computational linguistics, they are only beginning to be used in cognitive science. Since contextualized representations are closer to particular word senses, hence extra-linguistic categories, a promising approach might be to average the contextual representations of a noun, as used in a particular sense, from a number of sentences, to obtain a category representation. However, it is currently unknown to what extent contextualized embeddings can isolate a particular sense, and, in any case, as representations derived fundamentally from language, they are not expected to eliminate other aspects that disturb the mapping between language and the world, such as connotations. Moreover, standard distributional approaches can outperform contextualized representations in out-of-context semantic tasks (Lenci, Sahlgren, Jeuniaux, Gyllensten, & Miliani, 2021). These are important topics for a follow-up to the current paper, and for the growing body of work relying on contextualized embeddings in general.

Fourth, although many categories lack named instances, all categories do have instances to which one can refer. Accordingly, the NameBased approach could be extended by taking referring expressions other than names into account (and likewise for the NounBased model, for which one could use for instance multi-word expressions for non-lexicalized categories, e.g., "coffee table", "side table"). In particular, chains of co-referring expressions (e.g., "That maple", "the tree" and "it" in "That maple is huge. We planted the tree in 1980 and now it touches the roof.") exhibit the same referential stability that we argued benefits the NameBased model. For this reason a distributional representation of an entire coreference chain (as in Adel and Schütze 2014 or Clark and Manning 2016) could potentially supplant the use of names in an entity-based approach to category representation. While coreference resolution has been the focus of much research interest, it is an open question whether current models are good enough to build entity representations (Aina, Silberer, Sorodoc, Westera, & Boleda, 2019; Clark & Manning, 2015).

Finally, besides these concrete research directions, some of the central ideas in this paper carry over to other domains. One is the insight that certain parts of language are more rigidly tied to the world than others, and that this impacts the quality of distributional representations as models of concepts of extra-linguistic aspects. Another is the idea that one can get more reliable representations of category concepts by averaging representations of taxonomically subordinate notions (whether entities or subcategories). These ideas are not restricted to distributional semantics, as they can apply in principle to any field dealing with linguistically-induced concept representations. For instance, when extracting concept representations from neuro-imaging data one could consider using names as stimuli instead of nouns (e.g., for the noun/name distinction in ERP data see Proverbio, Mariani, Zani, and Adorni 2009). Our paper speaks to the complexity of the relationship between language and the world, the acknowledgment of which is paramount to enhancing our understanding of both natural language and human cognition.

Acknowledgments

We are grateful to Marco Marelli and two anonymous reviewers for their very helpful feedback, and to the associate editor Max Louwerse for his guidance. Abhijeet Gupta and Sebastian Padó received funding from Deutsche Forschungsgemeinschaft (DFG) through Sonderforschungsbereich 732, project D10. Gemma Boleda and Matthijs Westera received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154) and the Catalan government (SGR 2017 1575). Gemma Boleda also received funding from the Ramón y Cajal programme (grant RYC-2015-18907). Matthijs Westera also received funding from Leiden University (LUCL, SAILS). We gratefully acknowledge the support of NVIDIA Corporation with the donation of GPUs used for this research. This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains.

References

Abhishek, A. A., & Awekar, A. (2017). Fine-grained entity type classification by jointly learning representations and label embeddings. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics* (pp. 797–807). Valencia, Spain. Retrieved from `https://aclanthology.org/E17-1075`

Adel, H., & Schütze, H. (2014). Using mined coreference chains as a resource for a semantic task. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1447–1452). Doha, Qatar: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/D14-1151` doi: 10.3115/v1/D14-1151

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19–27). Boulder, Colorado. Retrieved from `https://www.aclweb.org/anthology/N09-1003`

Aina, L., Silberer, C., Sorodoc, I.-T., Westera, M., & Boleda, G. (2019). What do entity-centric models learn? Insights from entity linking in multi-party dialogue. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Lan-*

guage Technologies, volume 1 (long and short papers) (pp. 3772–3783). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/N19-1378` doi: 10.18653/v1/N19-1378

Alfonseca, E., & Manandhar, S. (2002). Distinguishing concepts and instances in WordNet. In *Proceedings of the First Global WordNet Conference.* Mysore, India.

Arora, S., Li, Y., Liang, Y., Ma, T., & Risteski, A. (2018). Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, *6*, 483–495. Retrieved from `https://transacl.org/ojs/index.php/tacl/article/view/1346/320`

Bach, K. (2002). Seemingly semantic intuitions. In J. Campbell, M. O'Rourke, & D. Shier (Eds.), *Meaning and truth.* New York: Seven Bridges Press.

Baldi, P., & Sandowski, P. (2013). Understanding dropout. In *Proceedings of Advances in Neural Information Processing Systems* (pp. 2814–2822).

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 238–247). Baltimore, Maryland: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/P14-1023` doi: 10.3115/v1/P14-1023

Baroni, M., & Lenci, A. (2011). How we BLESSed distributional semantic evaluation. In S. Padó & Y. Peirsman (Eds.), *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics* (pp. 1–10). Edinburgh, UK.

Baroni, M., Murphy, B., Barbu, E., & Poesio, M. (2010). Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, *34*(2), 222–254.

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. De Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 245–283). Oxford University Press.

Bianchi, F., & Palmonari, M. (2017). Joint learning of entity and type embeddings for analogical reasoning with entities. In *Nl4ai@ ai* ia* (pp. 57–68).

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. Retrieved from `https://aclanthology.org/Q17-1010.pdf` doi: 10.1162/tacl_a_00051

Boleda, G., Gupta, A., & Padó, S. (2017). Instances and concepts in distributional space. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 79–85). Valencia, Spain. Retrieved from `https://aclanthology.org/E17-2013`

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* (p. 1247–1250). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/1376616.1376746` doi: 10.1145/1376616.1376746

Borghi, A. M., & Binkofski, F. (2014). *Words as social tools: An embodied view on abstract concepts* (Vol. 2). Springer.

Bruni, E., Boleda, G., Baroni, M., & Tran, N. K. (2012). Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (pp. 136–145). Jeju Island, Korea. Retrieved from `https://aclanthology.org/P12-1015`

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1–47.

Casasola, M., Bhagwat, J., & Burke, A. S. (2009). Learning to form a spatial category of tight-fit relations: how experience with a label can give a boost. *Developmental psychology*, *45*(3), 711.

Chang, A., Spitkovsky, V. I., Manning, C. D., & Agirre, E. (2016). A comparison of named-entity disambiguation and word sense disambiguation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 860–867). Portorož, Slovenia: European Language Resources Association (ELRA). Retrieved from `https://aclanthology.org/L16-1139`

Clark, K., & Manning, C. D. (2015). Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1405–1415). Beijing, China: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/P15-1136` doi: 10.3115/v1/P15-1136

Clark, K., & Manning, C. D. (2016). Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 643–653). Berlin, Germany: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/P16-1061` doi: 10.18653/v1/P16-1061

Cruse, D. (1986). *Lexical semantics.* Cambridge University Press.

Curran, J. (2005). Supersense tagging of unknown nouns using semantic similarity. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (pp. 26–33). Ann Arbor, Michigan: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P05-1004` doi: 10.3115/1219840.1219844

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N19-1423` doi: 10.18653/v1/N19-1423

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database.* Cambridge, MA; London: The MIT Press.

Feng, Y., & Lapata, M. (2010). Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 91–99). Los Angeles, California: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N10-1011`

Făgărăşăn, L., Vecchi, E. M., & Clark, S. (2015, April). From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *Proceedings of the 11th International Conference on Computational Semantics* (pp. 52–57). London, UK: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/W15-0107`

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331.

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, *12*(4), 395–427. doi: 10.3758/BF03201693

Gouws, S., & Søgaard, A. (2015). Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1386–1390). Denver, Colorado: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N15-1157` doi: 10.3115/v1/N15-1157

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211.

Gupta, A., Boleda, G., Baroni, M., & Padó, S. (2015). Distributional vectors encode referential attributes.

In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 12–21). Lisbon, Portugal: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D15-1002` doi: 10.18653/v1/D15-1002

Guu, K., Miller, J., & Liang, P. (2015). Traversing knowledge graphs in vector space. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 318–327). Lisbon, Portugal: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D15-1038` doi: 10.18653/v1/D15-1038

Gärdenfors, P. (2000). *Conceptual spaces: the geometry of thought.* Cambridge, MA: MIT Press.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*. doi: 10.1177/1745691619861372

Hampton, J. (2015). Categories, prototypes and exemplars. In N. Riemer (Ed.), *The routledge handbook of semantics* (p. 124-141). London: Routledge.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335–346.

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

Herbelot, A. (2015). Mr Darcy and Mr Toad, Gentlemen: Distributional names and their kinds. In *Proceedings of the international conference on computational semantics* (pp. 151–161). Berlin, Germany.

Herbelot, A., & Baroni, M. (2017). High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 304–309). Copenhagen, Denmark. Retrieved from `https://www.aclweb.org/anthology/D17-1030`

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with genuine similarity estimation. *Computation Linguistics*, *41*(4), 665–695. doi: 10.1162/COLI_a_00237

Howell, D. C. (2012). *Statistical methods for psychology, eighth edition.* Cengage Learning.

Hutchinson, S., & Louwerse, M. (2018). Extracting social networks from language statistics. *Discourse Processes*, *55*(7), 607–618.

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

Jackendoff, R. (1990). *Semantic structures.* Cambridge, MA: The MIT Press.

Jones, M. N., Hills, T. T., & Todd, P. M. (2015). Hidden processes in structural representations: A reply to Abbott, Austerweil, and Griffiths (2015). *Psychological Review*, *122*(3), 570-–574.

Kelly, C., Devereux, B., & Korhonen, A. (2012). Semi-supervised learning for automatic conceptual property extraction. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)* (pp. 11–20). Montréal, Canada: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/W12-1702`

Kripke, S. (1980). *Naming and necessity.* Cambridge, MA: Harvard University Press.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology.* Beverly Hills, CA: Sage.

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

LaTourrette, A., & Waxman, S. R. (2019). A little labeling goes a long way: Semi-supervised learning in infancy. *Developmental Science*, *22*(1), e12736. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/desc.12736` doi: https://doi.org/10.1111/desc.12736

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, *20*(1), 1–31.

Lenci, A., Sahlgren, M., Jeuniaux, P., Gyllensten, A. C., & Miliani, M. (2021). A comprehensive comparative evaluation and analysis of distributional semantic models. *CoRR*, *abs/2105.09825*.

Retrieved from `https://arxiv.org/abs/2105.09825`

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225. Retrieved from `https://www.aclweb.org/anthology/Q15-1016` doi: 10.1162/tacl_a_00134

Levy, O., Remus, S., Biemann, C., & Dagan, I. (2015). Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 970–976). Denver, Colorado: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N15-1098` doi: 10.3115/v1/N15-1098

Ling, X., & Weld, D. S. (2012). Fine-grained entity recognition. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (p. 94–100). AAAI Press.

Louwerse, M. (2008). Embodied relations are encoded in language. *Psychonomic Bulletin & Review*, *15*(4), 838–844.

Louwerse, M. (2018). Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in cognitive science*, *10*(3), 573–589. doi: 10.1111/tops.12349

Louwerse, M., & Zwaan, R. (2009). Language encodes geographical information. *Cognitive Science*, *33*, 51–73. doi: 10.1111/j.1551-6709.2008.01003.x

Mandera, P., Keuleers, E., & Brysbaert, M. (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.

McCarthy, D., Koeling, R., Weeds, J., & Carroll, J. (2004). Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)* (pp. 279–286). Barcelona, Spain. Retrieved from `https://www.aclweb.org/anthology/P04-1036` doi: 10.3115/1218955.1218991

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th international conference on neural information processing systems - volume 2* (p. 3111–3119). Red Hook, NY, USA: Curran Associates Inc.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746–751). Atlanta: Association for Computational Linguistics. Retrieved from `http://www.aclweb.org/anthology/N13-1090`

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Five papers on WordNet. *International Journal of Lexicography*, *3*(4), 235–312.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and cognitive processes*, *6*(1), 1–28. doi: 10.1080/01690969108406936

Mitchell, J., & Lapata, M. (2010, nov). Composition in distributional models of semantics. *Cognitive science*, *34*(8), 1388–429. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/21564253` doi: 10.1111/j.1551-6709.2010.01106.x

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

Moreno, I., Romá-Ferri, M. T., & Moreda, P. (2017). Named entity classification based on profiles: A domain independent approach. In F. Frasincar, A. Ittoo, L. M. Nguyen, & E. Métais (Eds.), *Natural language processing and information systems* (pp. 142–146). Cham: Springer International Publishing.

Murphy, G. (2002). *The big book of concepts.* MIT Press.

Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of general psychology*, *135*(2), 151–166.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–61. doi: 10.1037/0096-3445.115.1.39

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N18-1202` doi: 10.18653/v1/N18-1202

Proverbio, A. M., Mariani, S., Zani, A., & Adorni, R. (2009). How are 'barack obama' and 'president elect' differentially stored in the brain? an erp investigation on the processing of proper and common noun pairs. *PloS one*, *4*(9). doi: 10.1371/journal.pone.0007126

Rigau, G., Atserias, J., & Agirre, E. (1997). Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of the 35th annual meeting of the association for computational linguistics* (pp. 48–55). Madrid, Spain. Retrieved from `http://www.aclweb.org/anthology/P97-1007` doi: 10.3115/976909.979624

Roller, S., & Erk, K. (2016). Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2163–2172). Austin, Texas: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D16-1234` doi: 10.18653/v1/D16-1234

Roller, S., Erk, K., & Boleda, G. (2014). Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 1025–1036). Dublin, Ireland: Dublin City University and Association for Computational Linguistics. Retrieved from `https://aclanthology.org/C14-1097`

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*, 192–233.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human perception and performance*, *2*(4), 491-502. doi: 10.1037/0096-1523.2.4.491

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Commun. ACM*, *8*(10), 627–633. Retrieved from `https://doi.org/10.1145/365628.365657` doi: 10.1145/365628.365657

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral & Brain Sciences*, *3*, 417-457.

Shimaoka, S., Stenetorp, P., Inui, K., & Riedel, S. (2017). Neural architectures for fine-grained entity type classification. In *Proceedings of the conference of the european chapter of the association for computational linguistics* (pp. 1271–1280). Valencia, Spain. Retrieved from `http://www.aclweb.org/anthology/E17-1119`

Shutova, E., Kaplan, J., Teufel, S., & Korhonen, A. (2013). A computational model of logical metonymy. *ACM Trans. Speech Lang. Process.*, *10*(3). Retrieved from `https://doi.org/10.1145/2483969.2483973` doi: 10.1145/2483969.2483973

Sikos, J., & Padó, S. (2019). Frame identification as categorization: Exemplars vs prototypes in embeddingland. In *Proceedings of the 13th international conference on computational semantics - long papers* (pp. 295–306). Gothenburg, Sweden: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/W19-0425` doi: 10.18653/v1/W19-0425

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Y. Bengio & Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015, conference track proceedings.* Retrieved from http://arxiv.org/abs/1409.1556

Smith, E., & Medin, D. (1981). *Categories and concepts.* Harvard University Press.

Smith, J. D. (2014). Prototypes, exemplars, and the natural history of categorization. *Psychonomic bulletin & review*, *21*(2), 312–331. doi: 10.3758/s13423-013-0506-0

Søgaard, A. (2016). Evaluating word embeddings with fMRI and eye-tracking. In *Proceedings of the 1st workshop on evaluating vector-space representations for NLP* (pp. 116–121). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W16-2521 doi: 10.18653/v1/W16-2521

Sommerauer, P., & Fokkens, A. (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* (pp. 276–286). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W18-5430 doi: 10.18653/v1/W18-5430

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological bulletin*, *87*(2), 245.

Toutanova, K., Chen, D., Pantel, P., Poon, H., Choudhury, P., & Gamon, M. (2015). Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1499–1509). Lisbon, Portugal: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D15-1174 doi: 10.18653/v1/D15-1174

Turney, P., & Litman, M. L. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning*, *60*(1-3), 251–278.

Vinner, S. (2002). The role of definitions in the teaching and learning of mathematics. In D. Tall (Ed.), *Advanced mathematical thinking* (Vol. 11, pp. 65–81). Dordrecht: Springer Netherlands. Retrieved from https://doi.org/10.1007/0-306-47203-1_5 doi: 10.1007/0-306-47203-1_5

Westbury, C. (2016). Pay no attention to that man behind the curtain: Explaining semantics without semantics. *The Mental Lexicon*, *11*(3), 350–374.

Westbury, C., & Hollis, G. (2019). Conceptualizing syntactic categories as semantic categories: Unifying part-of-speech identification and semantics using co-occurrence vector averaging. *Behavior Research Methods*, *51*(3), 1371–1398.

Westera, M., & Boleda, G. (2019). Don't blame distributional semantics if it can't do entailment. In *Proceedings of the 13th international conference on computational semantics - long papers* (pp. 120–133). Gothenburg, Sweden: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W19-0410 doi: 10.18653/v1/W19-0410

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th conference on natural language processing (konvens 2019): Long papers* (pp. 161–170). Erlangen, Germany: German Society for Computational Linguistics & Language Technology.

Xiao, M., & Guo, Y. (2014). Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the eighteenth conference on computational natural language learning* (pp. 119–129). Ann Arbor, Michigan: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W14-1613 doi: 10.3115/v1/W14-1613

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, *85*(3), 223-250. Retrieved from https://www.sciencedirect.com/science/article/pii/S0010027702001099 doi: https://doi.org/10.1016/S0010-0277(02)00109-9

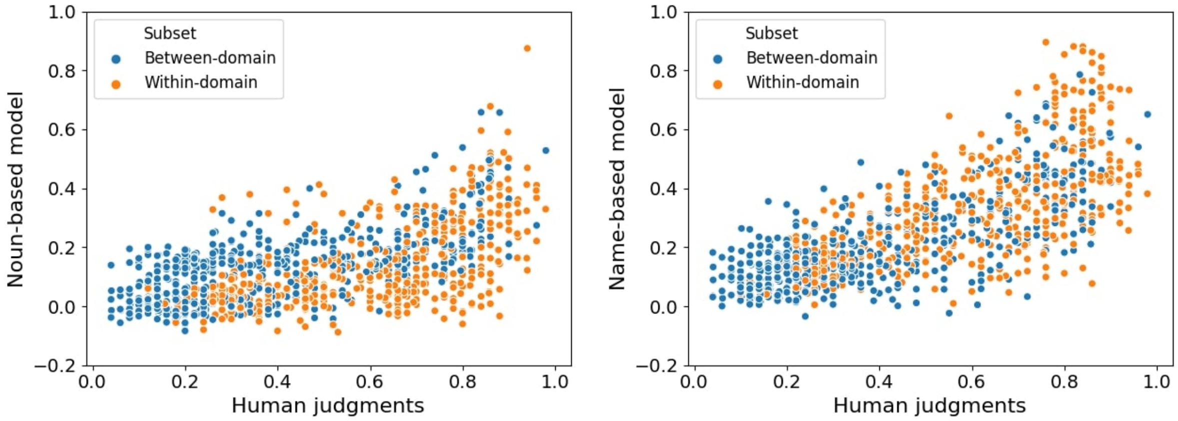


Figure 3: The NounBased (left) and NameBased (right) model similarity scores (y-axis) against human relatedness scores (x-axis).

Table 8: All data sets used for Experiment 2.

Subset	Number of pairs	of which positive	of which negative
Ent2Ent	9580	4790	4790
Inverse	9580	4790	4790
NotMemb-global	9580	4790	4790
NotMemb-inDomain	9580	4790	4790
Union-global	19160	4790	14370
Union-inDomain	19160	4790	14370

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *CoRR*, *abs/1212.5701*. Retrieved from `http://arxiv.org/abs/1212.5701`

A Appendix

A.1 Scatterplots for Experiment 1

Fig. 3 shows scatterplots for both models, where each point represents a pair of categories in terms of their human relatedness score (x-axis) and the respective model's prediction (y-axis); a distinction is drawn between within-domain and between-domain pairs.

A.2 Results for Subsets with a Single Type of Confounder

Recall from Section 4.2 that pairing confounders with the positive examples yields four different balanced subsets with one type of confounder each (Inverse, Ent2Ent, NotMemb-global and NotMemb-inDomain), along with the unions Union-global and Union-inDomain– see Table 8.

Tables 9 and 10 show the results for per-datapoint (micro) and per-category (macro) F_1-scores, respectively. They complement Table 5 in Section 4.2.

Recall that the datasets in Table 5 are arranged in their expected level of increasing difficulty from top to bottom, and indeed we see a clear trend of decreasing F_1-scores for both the NounBased and NameBased neural network models. As mentioned in the main text, the best NameBased model for a given subset is consistently better than the best NounBased network model. In the easiest subsets,

Table 9: Per-datapoint (micro F_1-scores) results on the subsets with one single type of confounder.

			NounBased			NameBased		
Dataset	BL_{Freq}	BL_{Pos}	Cos	NN-1HL	NN-2HL	Cos	NN-1HL	NN-2HL
Inverse	0.50	0.67	0.67	0.98	0.98	0.67	**0.99**	0.94
Ent2Ent	0.50	0.67	0.67	0.90	0.91	0.86	**0.92**	0.86
NotMemb-global	0.50	0.67	0.71	0.85	0.84	0.88	**0.90**	0.89
NotMemb-inDomain	0.50	0.67	0.69	0.72	0.67	**0.78**	**0.78**	0.74
Union-global	0.25	0.40	0.43	0.74	0.77	0.59	**0.85**	0.75
Union-inDomain	0.25	0.40	0.41	0.51	0.65	0.55	**0.76**	0.68

Table 10: Per-category (macro F_1-scores) results on the subsets with one single type of confounder.

			NounBased			NameBased		
Dataset	BL_{Freq}	BL_{Pos}	Cos	NN-1HL	NN-2HL	Cos	NN-1HL	NN-2HL
Inverse	0.42	0.67	0.67	0.94	0.93	0.67	**0.96**	0.90
Ent2Ent *	0.51	1.00*	1.00*	0.68	0.74	0.86	**0.97**	0.87
NotMemb-global	0.35	0.59	0.60	0.65	0.65	0.69	**0.73**	**0.73**
NotMemb-inDomain	0.38	0.59	0.60	0.54	0.55	0.61	**0.64**	0.63
Union-global *	0.20	0.43	0.41	0.38	0.45	0.45	**0.65**	0.53
Union-inDomain *	0.21	0.43	0.40	0.21	0.37	0.42	**0.59**	0.47

*: The Ent2Ent items of these datasets do not contain any category, hence do not affect macro per-category F_1 score – hence the perfect score of the positive baseline and NounBased Cos on Ent2Ent.

Inverse and Ent2Ent, the improvement is in general minor; as already shown in Boleda et al. (2017) standard word embeddings contain information to draw the high-level distinction between categories and entities, and this is all that is needed for these subsets.[3] However, on the more challenging NotMemb subsets, the difference increases to 6-11 points.

Table 5 also shows that the neural network models beat their respective cosine models Cos in most subsets, though occasionally only by a small margin, and with an exception in NotMemb-inDomain where the NameBased cosine model and the one-layer neural network NN-1HL share first place (0.78).

A.3 Analysis of Cosine Models

In addition to supporting our hypothesis that NameBased representations are better, the cosine models mostly show an expected pattern – no better than baseline on Inverse (as cosine is a symmetric measure, thus unable to distinguish a pair from its inverse) and increasing difficulty from NotMemb-global downwards – with a noteworthy contrast in Ent2Ent, where the NounBased cosine model performs no better than the positive baseline while the NameBased cosine model reaches as high as 0.86. The latter is noteworthy because one might have expected the NameBased representations to be at a disadvantage here: Ent2Ent tests a model's ability to distinguish categories from entities, and since NameBased representations model a category concept as the average of its entity name vectors, one might have expected these representations to be more 'entity-like' than a corresponding NounBased vector, hence more conducive to confusing entities and categories. But the NameBased cosine model score on Ent2Ent shows that this is not the case.

To understand this pattern, Figure 4 shows the distributions of cosine similarities (i) of the entity-

[3] The slight edge for NameBased models on these two datasets is noteworthy (as for the cosine model on Ent2Ent discussed below), because one might have expected the NameBased representations to be more entity-like and hence be at a disadvantage here.

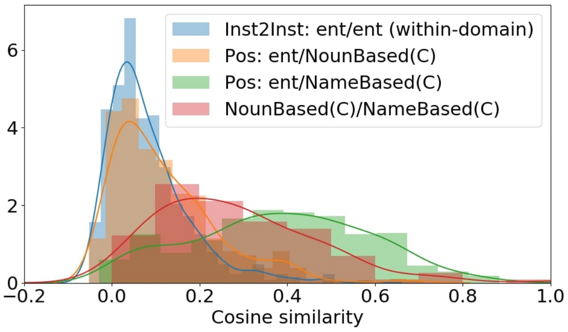


Figure 4: Distribution of cosine similarities on the positive and confounder items of Ent2Ent (test set), and of cosine similarities between NounBased(C) and NameBased(C) representations of the same category C.

entity confounder pairs in Ent2Ent in blue, (ii) of the true, entity-category pairs in Pos according to the NounBased model in orange, and (iii) of the same pairs according to the NameBased model in green – the fourth, red curve is explained shortly.[4] The figure shows that the cosines for entity-entity confounders (blue) show a very similar distribution to the cosines for true pairs according to the noun-based model (orange), but are substantially different from the cosines for true pairs according to the name-based model (green). As a consequence, for the NounBased model no cosine threshold exists that can really separate entities from categories, and the model performs at baseline level (F_1-score 0.67). By contrast, for the NameBased model a reasonable threshold does exist (around 0.2 on the x-axis), and the model performs much better (F_1-score 0.86).

As an interesting aside, the fourth, red curve in Figure 4 represents the distribution of pairs of NounBased(C) and NameBased(C) representations of the same category C. It shows that the two representations of the same category are quite similar to each other (red), indeed, more so than individual entities and their NounBased categories (orange). To phrase this more intuitively: by taking a name vector and averaging it with other name vectors, the representation becomes more noun-like (Herbelot, 2015, Fig. 2, showed this qualitatively). Together with the high similarity of the NameBased representation to its individual entities (green curve), this suggests that NameBased representations can be seen, interestingly, as a kind of midway point between names and nouns (i.e., red lies between orange and green).

[4] The plotted distributions have been restricted to categories and entities in the test set of Ent2Ent, to ensure that the NameBased representations are not built from the entities to which we compare them (see Section 4.2).