# Logical features in distributional word representations

**Tal Linzen**[1,2] **Emmanuel Dupoux**[1] **Benjamin Spector**[2]
[1]Laboratoire de Sciences Cognitives et Psycholinguistique [2]Institut Jean Nicod
École Normale Supérieure
PSL Research University
{tal.linzen, benjamin.spector}@ens.fr
emmanuel.dupoux@gmail.com

## Abstract

Do distributional word representations encode the linguistic regularities that theories of meaning argue they should encode? We address this question in the case of the logical properties (monotonicity, force) of quantificational words such as *everything* and *always*. Using the vector offset approach to solving word analogies, we find that the skip-gram model of distributional semantics behaves in a way that is consistent with encoding these features in some areas, but not in others. Human participants performed well even where the model struggled. The model's success crucially depended on large training corpora, suggesting that distributional information is insufficient for human language acquisition. Finally, we discuss caveats with using the offset method to uncover the representation of linguistic features.

## 1 Introduction

Vector-space models of lexical semantics (VSMs) represent words as points in a high-dimensional space. Similar words are represented by points that are close together in the space. VSMs are typically trained on a corpus in an unsupervised way; the goal is for words that occur in similar contexts to be assigned similar representations. The context of a word in a corpus is often defined as the set of words that occur in a small window around the word of interest (Lund and Burgess, 1996; Turney and Pantel, 2010). VSM representations have been shown to be useful in improving the performance of NLP systems (Turian et al., 2010; Bansal et al., 2014) as well as in predicting cognitive measures such as similarity judgments and semantic priming (Jones et al., 2006; Hill et al., 2015).

While there is evidence that VSM representations encode useful information about the meaning of open-class words such as *dog* or *table*, less is know about the extent to which they capture abstract linguistic properties, in particular the aspects of word meaning that are crucial in logical reasoning. Some have conjectured that those properties are unlikely to be encoded in VSMs (Lewis and Steedman, 2013), but evidence that VSMs encode features such as syntactic category or verb tense suggests that this pessimism is premature (Mikolov et al., 2013c; Levy and Goldberg, 2014).

The goal of this paper is to evaluate to what extent logical features are encoded in VSMs. We undertake a detailed analysis of words with quantificational features, such as *everybody* or *nowhere*. To assess whether a particular linguistic feature is encoded in a vector space, we adopt the vector offset approach to the analogy task (Turney, 2006; Mikolov et al., 2013c; Dunbar et al., 2015). In the analogy task, a system is requested to fill in the blank in a sentence:

(1) *man* is to *woman* as *king* is to ___.

The system is expected to infer the relation between the first two words—*man* and *woman*—and find a word that stands in the same relation to *king*. When this taks is solved using the offset method, there is no explicit set of relations that the system is trained to identify. We simply subtract the vector for *man* from the vector for *woman* and add it to *king*. If the offset *woman* − *man* represents an abstract gender feature, adding that offset to *king* should lead us to *queen* (Fig. 1).

If purely distributional representations successfully encode linguistic features and perform linguistic tasks as well as humans do, algorithms that rely on other sources of information may not be

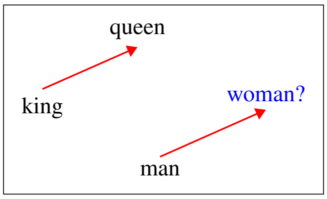


Figure 1: Using the vector offset method to solve the analogy task (Mikolov et al., 2013c).

necessary to account for human learning (Redington et al., 1998). These cognitive considerations lead us to conduct two additional experiments. First, we evaluate how well humans solve the analogy task that we expect our VSMs to solve. Second, we investigate how the quality of the representations degrades as the size of the training corpus approaches the amount of words that a child is likely to be exposed to when learning a language.

An additional contribution of our study is methodological. Although the offset method has become a part of the battery of tests used to compare VSMs, its properties are not always explored in great detail. We focus in particular on understanding the extent to which this method reflects the encoding of a linguistic feature in the geometry of the space.

A large and constantly expanding range of VSM architectures have been proposed in the literature (Mikolov et al., 2013a; Pennington et al., 2014; Turney and Pantel, 2010). Instead of exploring the full range of architectures, the present study will focus on the skip-gram model, implemented in `word2vec` (Mikolov et al., 2013b). This model has been argued to perform either better than or on a par with competing architectures, depending on the task and on hyperparameter settings (Baroni et al., 2014; Levy et al., 2015). Particularly pertinent to our purposes, Levy et al. (2015) find that the skip-gram model tends to recover formal linguistic features more accurately than traditional distributional models.

# 2 Quantificational words

We focus on words that quantify over the elements of a domain, such as *everyone* or *nowhere*. We restrict our attention to single words that include the domain of quantification as part of their meaning – that is, we exclude determiners (*every*) and phrases (*every person*). The meaning of a quantifier is determined by three factors: quantificational force, polarity and domain of quantification. We describe these factors in turn.

## 2.1 Quantificational force

We focus on universal and existential quantificational words, which can be translated into first-order logic using a universal ($\forall$) or existential ($\exists$) quantifier. For example, *everybody* and *nobody* are both universal:

(2) Everybody smiles:
$\forall x.\mathit{person}(x) \rightarrow \mathit{smiles}(x)$

(3) Nobody smiles:
$\forall x.\mathit{person}(x) \rightarrow \neg\mathit{smiles}(x)$

*Somebody* is existential:

(4) Somebody smiles:
$\exists x.\mathit{person}(x) \wedge \mathit{smiles}(x)$

English has quantificational expressions that don't fall into either category (*three people*, *most things*). Those are usually not encoded as a single English word, and are therefore not considered in this paper.

## 2.2 Polarity

Quantifiers that can be expressed as a single word are in general either increasing or decreasing. A quantifier is increasing if any predicate that is true of the quantifier can be broadened without affecting the truth value of the sentence (Barwise and Cooper, 1981). For example, since *everyone* is increasing, (5-a) entails (5-b):

(5) a. Everybody went out to a death metal concert last night.
b. Everybody went out last night.

By contrast, in decreasing quantifiers such as *nobody* the truth of broader predicates entails the truth of narrower ones:

(6) a. Nobody went out last night.
b. Nobody went out to a death metal concert last night.

## 2.3 Domain

We studied six domains. The first three domains are intuitively straightforward: PERSON (e.g., *everybody*); OBJECT (e.g., *everything*); and PLACE (e.g., *everywhere*). The three additional domains are described below.

|  | INC. |  | DEC. |
|---|---|---|---|
|  | Universal | Existential | Universal |
| PERSON | *everybody* | *somebody* | *nobody* |
| OBJECT | *everything* | *something* | *nothing* |
| PLACE | *everywhere* | *somewhere* | *nowhere* |
| TIME | *always* | *sometimes* | *never* |
| MODAL | *must* | *can* | *cannot* |
| MODAL V. | *require* | *allow* | *forbid* |

Table 1: All of the words tested in the experiments (INC = Increasing, DEC = Decreasing).

**TIME:** Temporal adverbs such as *always* and *seldom* are naturally analyzed as quantifying over situations or events (Lewis, 1975; de Swart, 1993). The sentence *Caesar always awoke before dawn*, for example, can be seen as quantifying over waking events and stating that each of those events occurred before dawn.

**MODAL:** Modal auxiliaries such as *must* or *can* quantify over relevant possible worlds (Kripke, 1959). Consider, for example, the following sentences:

(7) a. Anne must go to bed early.
b. Anne can go to bed early.

Assuming deontic modality, such as the statement of a rule, (7-a) means that in all worlds in which the rule is obeyed, Anne goes to bed early, whereas (7-b) means that there exists at least one world consistent with the speaker's orders in which she goes to bed early.

**MODAL VERB:** Verbs such as *request* and *forbid* can be paraphrased using modal auxiliaries: *he allowed me to stay up late* is similar in meaning to *he said I can stay up late*. It is plausible to argue that *allow* is existential and increasing, just like *can*.

## 3 Evaluation

In what follows, we use the following notation (Levy and Goldberg, 2014):

(8) $a : a^* :: b :$ ___

The offset model is typically understood as in Figure 1: the analogy task is solved by finding $x = a^* - a + b$. In practice, since the space is continuous, $x$ is unlikely to precisely identify a word in the vocabulary. The guess is then taken to be the word $x^*$ that is nearest to $x$:

$$x^* = \arg\max_{x'} \cos(x', a^* - a + b) \qquad (1)$$

where cos denotes the cosine similarity between the vectors. This point has a significant effect on the results of the offset method, as we will see below. Following Mikolov et al. (2013c) and Levy and Goldberg (2014), we normalize $a$, $a^*$ and $b$ prior to entering them into Equation 1.

**Trivial responses:** $x^*$ as defined above is almost always trivial: in our experiments the nearest neighbor of $x$ was either $a^*$ (11% of the time) or $b$ (88.9% of the time). Only in a single analogy out of the 2160 we tested was it not one of those two options. Following Mikolov et al. (2013c), then, our guess $x^*$ will be the nearest neighbor of $x$ that is not $a$, $a^*$ or $b$.

**Baseline:** The fact that the nearest neighbor of $a^* - a + b$ tends to be $b$ itself suggests that $a^* - a$ is typically small in comparison to the distance between $b$ and any of its neighbors. Even if $b$ is excluded as a guess, then, one might be concerned that the analogy target $b^*$ is closer to $b$ than any of its neighbors. If that is the case, our success on the analogy task would not be informative: our results would stay largely the same if $a^* - a$ were replaced by a random vector of the same magnitude. To address this concern, we add a baseline that solves the analogy task by simply returning the nearest neighbor of $b$, ignoring $a$ and $a^*$ altogether.

**Multiplication:** Levy and Goldberg (2014) point out that the word $x^*$ that is closest to $a^* - a + b$ in terms of cosine similarity is the one that maximizes the following expression:

$$\arg\max_{x'} (\cos(x', a^*) - \cos(x', a) + \cos(x', b)) \qquad (2)$$

They report that replacing addition with multiplication improves accuracy on the analogy task:

$$\arg\max_{x'} \frac{\cos(x', a^*) \cos(x', b)}{\cos(x', a)} \qquad (3)$$

We experiment with both methods.

**Synonyms:** Previous studies required an exact match between the guess and the analogy target selected by the experimenter. This requirement may underestimate the extent to which the space encodes linguistic features, since the bundle of semantic features expressed by the intended target can often be expressed by one or more other words. This is the case for *everyone* and *everybody*, *prohibit* and *forbid* or *can't* and *cannot*. As such, we considered synonyms of $b^*$ to be exact matches. Likewise, we considered synonyms of $a$, $a^*$ and $b$ to be trivial responses and excluded them from consideration as guesses.

This treatment of synonyms is reasonable when the goal is to probe the VSM's semantic representations (as it often is), but may be inappropriate for other purposes. If, for example, the analogy task is used as a method for generating inflected forms, *prohibiting* would not be an appropriate guess for *like : liking :: forbid : ___*.

**Partial success metrics:** We did not restrict the guesses to words with quantificational features: all of the words in the vocabulary, including words like *penguin* and *melancholy*, were potential guesses. In addition to counting exact matches ($x^* = b^*$), then, we keep track of the proportion of cases in which $x^*$ was a quantificational word in one of the six relevant domains.

Within the cases in which $x^*$ was a quantificational word, we separately counted how often $x^*$ had the expected domain, the expected polarity and the expected force. To be able to detect such partial matches, we manually added some words to our vocabulary that were not included in the set in Table 1. These included items starting with *any*, such as *anywhere* or *anybody*, as well as additional temporal adverbs (*seldom*, *often*).

Finally, we record the rank of $b^*$ among the 100 nearest neighbors of $x$, where a rank of 1 indicates an exact match. It was often the case that $b^*$ was not among the 100 nearest neighbors of $x$; we therefore record how often $b^*$ was ranked at all.

## 4 Experimental setup

### 4.1 Analogies

For each ordered pair of domains ($6 \times 5 = 30$ pairs in total), we constructed all possible analogies where $a$ and $a^*$ were drawn from one domain (the source domain) and $b$ and $b^*$ from the other (the target domain). Since there are three words per domain, we had six possible analogies per domain pair, for a total of 180 analogies.

Each set of four words was used to construct multiple analogies. Those analogies are in general not equivalent. For example, the words *everybody*, *nobody*, *everywhere* and *nowhere* make up the following analogies:

(9) *everybody : nobody :: everywhere : ___*

(10) *nobody : everybody :: nowhere : ___*

(11) *everywhere : nowhere :: everybody : ___*

(12) *nowhere : everywhere :: nobody : ___*

The neighborhoods of *everywhere* and *nobody* may well differ in density. Since the density of the neighborhood of $b$ affects the results of the offset method, the result is not invariant to a permutation of the words in an analogy. It is, however, invariant to replacing a within-domain analogy with an across-domain one. The following analogy is equivalent to (9):

(13) *everybody : everywhere :: nobody : ___*

This analogy would be solved by finding the nearest neighbor of *everywhere* − *everybody* + *nobody*, which is, of course, the same as the nearest neighbor of *nobody* − *everybody* + *everywhere* used to solve (9). We do not include such analogies.

### 4.2 VSMs

We trained our VSMs using the skip-gram with negative sampling algorithm implemented in `hyperwords`,[1] which extends `word2vec` to allow finer control over hyperparameters. The vectors were trained on a concatenation of ukWaC (Baroni et al., 2009) and a 2013 dump of the English Wikipedia, 3.4 billion words in total.

The skip-gram model has a large number of parameters. We set most of those parameters to values that have been previously shown to be effective (Levy et al., 2015); we list those values below. We only vary three parameters that control the context window. Syntactic category information has been shown to be best captured by narrow context windows that encode the position of the context word relative to the focus word (Redington et al., 1998; Sahlgren, 2006). Our goal in varying these parameters is to identity the contexts that are most conducive to recovering logical information.

**Window size:** We experimented with context windows of 2, 5 or 10 words on either side of the focus word (i.e., a window of size 2 around the focus word consists of four context words).

**Window type:** When constructing the vector space, the skip-gram model performs frequency-based pruning: rare words are discarded in all

[1] https://bitbucket.org/omerlevy/hyperwords

cases and very frequent words are discarded probabilisitically. We experimented with static and dynamic windows. The size of static windows is determined prior to frequency-based word deletion. By contrast, the size of dynamic windows is determined after frequent and infrequent words are deleted. This means that dynamic windows often include words that are farther away from the focus words than the nominal window size, and that words that tend to have very frequent function words around them will systematically have a larger effective context window.

**Context type:** We experimented with bag-of-words (nonpositional) contexts and positional contexts. In nonpositional contexts, a context word *cat* is treated in the same way regardless of its distance from the focus word and of whether it follows or precedes it. In positional contexts, on the other hand, context words are annotated with their position relative to the focus words; the context word $cat^{-2}$ is considered to be distinct from $cat^{+1}$.

**Fixed hyperparameters:** We used the following values for the rest of the hyperparameters: 500-dimensional words vectors; 15 negative samples per focus word; words with a frequency of less than 100 were discarded; words with unigram probability above $10^{-5}$ were probabilistically discarded (preliminary experiments showed that a $10^{-3}$ threshold reduced performance across the board); negative samples were drawn from the unigram frequency distribution, after that distribution was smoothed with exponent $\alpha = 0.75$; we performed one iteration through the data.

## 5 Results

**Offset method:** Overall accuracy was fairly low (mean: 0.29, range: $0.23 - 0.35$), somewhat lower than the 0.4 accuracy that Mikolov et al. (2013c) report for their syntactic features. Strikingly, $b^*$ was among the 100 nearest neighbors of $x$ only in 70% of the cases. When the guess was a quantificational word (61% of the time), it was generally in the right domain (93%). Its polarity was correct 72% of the time, and its force 54% of the time.

The static nonpositional 5-word VSM achieved the best accuracy (35%), best average rank (5.5) and was able to recover the most quantificational features (polarity: 82% correct; force: 63% correct; both proportions are conditioned on the guess being a quantificational word).

| Size | Context | Window | B | O | M | O - B |
|---|---|---|---|---|---|---|
| 2 | Nonpos | Dynamic | .08 | .32 | .34 | .24 |
| 2 | Nonpos | Static | .06 | .23 | .24 | .17 |
| 2 | Pos | Dynamic | .06 | .29 | .32 | .24 |
| 2 | Pos | Static | .06 | .24 | .27 | .19 |
| 5 | Nonpos | Dynamic | .07 | .28 | .29 | .22 |
| 5 | Nonpos | Static | .11 | .35 | .36 | .24 |
| 5 | Pos | Dynamic | .03 | .29 | .31 | .27 |
| 5 | Pos | Static | .06 | .28 | .29 | .23 |
| 10 | Nonpos | Dynamic | .08 | .28 | .29 | .19 |
| 10 | Nonpos | Static | .17 | .31 | .31 | .14 |
| 10 | Pos | Dynamic | .17 | .32 | .31 | .16 |
| 10 | Pos | Static | .11 | .26 | .26 | .15 |

Table 2: Results on all hyperparameter settings, evaluated using three methods: B(aseline), O(ffset) and M(ultiplication).

**Alternatives to the offset method:** In line with the results reported by Levy and Goldberg (2014), we found that substituting multiplication for addition resulted in slightly improved performance in 10 out of 12 VSMs, though the improvement in each individual VSM was never significant according to Fisher's exact test (Table 2). If we take each VSM to be an independent observation, the difference across all VSMs is statistically significant in a t-test ($t = 2.45, p = 0.03$).

The baseline that ignores $a$ and $a^*$ altogether reached an accuracy of up to 0.17, sometimes accounting for more than half the accuracy of the offset method. The success of the baseline is significant, given that chance level is very low (recall that all but the rarest words in the corpus were possible guesses). Still, the offset method was significantly more accurate than the baseline in all VSMs ($10^{-12} < p < 0.003$, Fisher's exact test).

**Differences across domains:** We examine the performance of the offset method in the best-performing VSM in greater detail. There were dramatic differences in accuracy across target domains. When $b^*$ was a PERSON, guesses were correct 73% of the time; the correct guess was one of the top 100 neighbors 87% of the time, and its average rank was 1.31. Conversely, when $b^*$ was a MODAL VERB, the guess was never correct; in fact, $b^*$ was one of the 100 nearest neighbors of $x$ only 7% of the time, and the average rank in these cases was 59. Variability across source domains was somewhat less pronounced. Fig. 2a shows the interaction between source and target domain.

In light of the differences across domains, we repeated our investigation of the influence of context parameters, this time restricting the source and target domains to PERSON, PLACE and OBJECT. Exact match accuracy ranged from 0.5 for the static nonpositional 2-word window to 0.83 for the static nonpositional 5-word window. The latter VSM achieved almost perfect accuracy in cases where the guess was a quantificational word (domain: 1.0, polarity: 0.97, force: 1.0). We conclude that in some domains logical features can be robustly recovered from distributional information.

**Effect of context parameters:** Overall, the influence of context parameters on accuracy was not dramatic. When the VSMs are compared based on the extent that the offset method improves over the baseline (O - B in Table 2), a somewhat clearer picture emerges: the improvement is greatest in intermediate window sizes, either 5-word windows or dynamic 2-word windows. This contrasts with findings on the acquisition of syntactic categories, where narrower contexts performed best (Redington et al., 1998), suggesting that the cues to quantificational features are further from the focus word than cues to syntactic category.

One candidate for such a cue is the word's compatibility with negative polarity items (NPI) such as *any*. NPIs are often licensed by decreasing quantifiers (Fauconnier, 1975): *nobody ate any cheese* is grammatical, but **everybody ate any cheese* isn't. Whereas contextual cues to syntactic category—e.g., *the* before nouns—are often directly adjacent to the focus word, *any* will typically be part of a different constituent from the focus word, and is therefore quite likely to fall outside a narrow context window.

We did not find a systematic effect of the type of context (positional vs. nonpositional). However, as Section 7 below shows, this parameter does affect performance when the VSMs are trained on smaller corpora.

## 6 How well do humans do the task?

Some of the analogies are intuitively fairly difficult: quantification over possible deontic worlds (*require* vs. *forbid*) is quite different from quantification over individuals (*everybody* vs. *nobody*). Those are precisely the domains in which the VSMs performed poorly. Are we asking too much of our VSM representations? Can humans perform this task?[2]

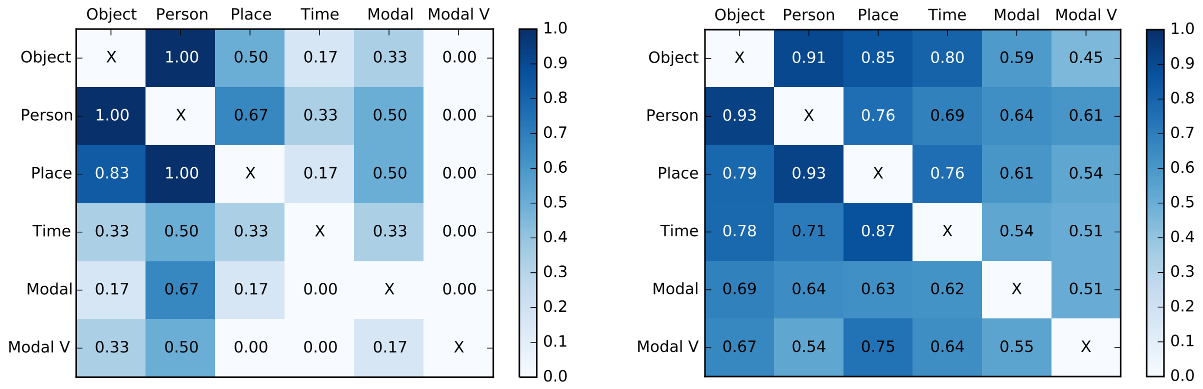
To answer this question, we gave the same analogies to human participants recruited through Amazon Mechanical Turk. We divided our 180 quantificational analogies into five lists of 36 analogies each. Each list additionally contained four practice trials presented in the beginning of the list and ten catch trials interspersed throughout the list. These additional trials contained simple analogies, such as *big : bigger :: strong :* ___ or *brother : sister :: son :* ___. Each of the lists was presented to ten participants (50 participants in total). They were asked to type in a word that had the same relationship to the third word as the first two words had to each other.

We excluded participants that made more than three mistakes on the catch trials (three participants) as well as one participant who did not provide any answer to some of the questions. While mean accuracy varied greatly among subjects (range: $0.22 - 1$; mean: $0.68$; median: $0.69$; standard deviation: $0.17$), it was in general much higher than the accuracy of the VSMs.

Fig. 2b presents the human participants' average accuracy by source and target domain. Mean accuracy was 0.45 or higher for all combinations of source and target domains. Logistic regression confirmed that having MODAL VERB and MODAL as either the source or target domain led to lower accuracy. There were no statistically significant differences between those two domains or among the remaining four domains, with the exception of TIME as a target domain, which was less accurate than PLACE, OBJECT and PERSON.

The VSMs did not have access to the morphological structure of the words. This makes the comparison with humans difficult: it is hard to see how human participants could be stopped from accessing that knowledge when performing an analogy such as *nowhere : somewhere :: nobody :* ___. Notably, however, the difference in performance between the morphologically marked domains and the

[2] These two questions are highly related from our cognitive modeling perspective, but in general it is far from clear that human performance on a logical task is an appropriate yardstick for a computational reasoning system. In the domain of quantifier monotonicity, in particular, there are documented discrepancies between normative logic and human reasoning (Chemla et al., 2011; Geurts and van Der Slik, 2005). In many cases it may be preferable for a reasoning system to conform to normative logic rather than mimic human behavior precisely.



(a) VSM: offset method

(b) Human responses

Figure 2: On the left: accuracy of the best model (static nonpositional 5-word context), broken down by source (in the y-axis) and target (in the x-axis) domain. On the right: human responses.

other domains is if anything *more* marked in the VSMs than in humans. Moreover, there is a fairly small difference in the accuracy of our human participants between PLACE and TIME as target domains, even though the former is morphologically marked and the latter isn't.

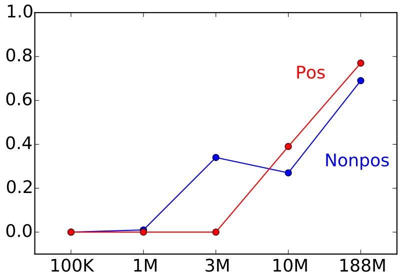


Figure 3: Effect of training corpus size on the accuracy of the analogy task, averaged across vector size and window size.

## 7 Cognitively plausible corpus sizes

Distributional information is crucial for learning the meaning of words such as quantificational words that do not have a clear correlate in visual experience (Gleitman et al., 2005). Is the information captured by the VSMs we have considered sufficient for acquiring quantificational words under cognitively plausible circumstances? The ideal testing ground for this question would be infant directed speech corpora. As a first step, however, we investigate how the VSMs perform when trained on subsets of our training corpus that contain an amount of data that a human might encounter when learning a language.

Hart and Risley (1995) estimate that American children are exposed to between 3 and 11 million words every year, depending on the socioeconomic status of their family. We sampled four subcorpora from our Wikipedia corpus, with 100K, 1M, 3M and 10M sentences. As the average sentence length in the corpus is 18 words, the corpora contained 1.8M, 18M, 54M and 180M tokens, respectively. The 1M and 3M sentence corpora represent plausible amounts of exposure for a child.

Given that VSM accuracy was low in some of the domains even when the spaces were trained on 3.4G tokens, we limit our experiments in this section to the OBJECT and PERSON domains. We made two changes to the hyperparameters settings that were not modulated in the VSMs trained on the full corpus. First, we lowered the threshold for rare word deletion (100K / 1M sentences: 10; 3M sentences: 50; 10M sentences: 100). Second, we experimented with smaller vectors (100, 300 and 500), under the assumption that it may be more difficult to train large vectors on a small data set. We again experimented with window sizes of 2, 5 and 10 words on either side of the focus word and with positional and nonpositional contexts. The size of the windows was always static.

Fig. 3 shows the accuracy of the analogy task averaged across vector sizes and window sizes. VSMs trained on the 100K and 1M subcorpora completely failed to perform the task: with the exception of one model that performed one out the 12 analogies correctly, accuracy was always 0. The VSMs trained on the 3M and 10M sentences subcorpora perform better (between 0.27 and 0.39 on average), though still much worse than the VSMs trained on the full corpus. The type of context had a large effect on the success of the model: positional contexts trained on the 3M subcorpus completely failed to do the task, whereas on the 10M subcorpus they performed better than nonpositional ones. The performance advantage of positional contexts was larger on the 10M corpus than on the full corpus.

The degraded performance of the VSMs on smaller training corpora suggests that bag-of-words distributional information alone is unlikely to be sufficient for the acquisition of quantification. An adequate cognitive model would need to consider richer types of context, such as syntactic context and discourse structure, or to make explicit reference to the way these words are used in logical reasoning.

## 8 Related work

There is a large body of work on the evaluation of VSMs (Turney and Pantel, 2010; Hill et al., 2015). A handful of recent papers have looked at distributional representations of logical words. Baroni et al. (2012) extracted corpus-based distributional representations for quantifier phrases such as *all cats* and *no dogs*, and trained a classifier to detect entailment relations between those phrases; for example, the classifier might learn that *all cats* entails *some cats*. Bernardi et al. (2013) introduce a phrase similarity challenge that relies on the correct interpretation of determiners (e.g., *orchestra* is expected to be similar to *many musicians*), and use it to evaluate VSMs and composition methods. Hermann et al. (2013) discuss the difficulty of accounting for negation in a distributional semantics framework.

Another line of work seeks to combine the graded representations of content words such as *mammal* or *book* with a symbolic representation of logical words (Garrette et al., 2014; Lewis and Steedman, 2013; Herbelot and Vecchi, 2015). Our work, which focuses on the quality of graded representation of logical words, can be seen as largely orthogonal to this line of work.

Finally, our study is related to recent neural network architectures designed to recognize entailment and other logical relationships between sentences (Bowman et al., 2014; Rocktäschel et al., 2015). Those systems learn word vector representations that are optimized to perform an explicit entailment task (when trained in conjunction with a compositional component). In future work, it may be fruitful to investigate whether those representations encode logical features more faithfully than the unsupervised representations we experimented with.

## 9 Conclusion

The skip-gram model, like earlier models of distributional semantics, represents words in a vector space using only their bag-of-words contexts in a corpus. We tested whether the representations that this model acquires for words with quantificational content encode the logical features that theories of meaning predict they should encode. We addressed this question using the offset method for solving the analogy task, $a : a^* :: b :$ ___ (e.g., *everyone* : *someone* :: *everywhere* : ___).

We made several methodological observations regarding this method, expanding on Levy and Goldberg (2014). First, when implemented literally the guess it provides is almost always trivial (either $a^*$ or $b$), casting doubt on a strong geometric interpretation of its results. Second, requiring an all-or-nothing match with an intended analogy target is too stringent: this evaluation method assesses the encoding of multiple semantic features at once, and doesn't take into account the fact that a bundle of semantic features can be expressed in multiple ways (e.g., *can't* and *cannot*). Most importantly, given the central role of cosine similarity in the offset method, this method should be evaluated relative to a baseline that only takes similarity to $b$ into account, ignoring $a$ and $a^*$.

These issues aside, we showed that distributional methods successfully recovered quantificational features. Accuracy was higher when the context window was of an intermediate size, sometimes approaching 100% on simpler domains. Performance on other domains was poorer, however. Humans given the same task also showed variability across domains, but achieved better accuracy overall, suggesting that there is room for improving the VSM representations. Finally, accuracy dropped dramatically when training corpus size approached the amount of words that a child might be exposed to when learning the language. This suggests that bag-of-words distributional methods are inefficient, and that human learning of quantificational features relies on additional, more structured sources of information.

## References

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 809–815, Baltimore, Maryland, June. Association for Computational Linguistics.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 23–32. Association for Computational Linguistics.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 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*, pages 238–247.

Jon Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4(2):159–219.

Raffaella Bernardi, Georgiana Dinu, Marco Marelli, and Marco Baroni. 2013. A relatedness benchmark to test the role of determiners in compositional distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 53–57. Association for Computational Linguistics.

Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. 2014. Recursive neural networks can learn logical semantics. *arXiv preprint arXiv:1406.1827*.

Emmanuel Chemla, Vincent Homer, and Daniel Rothschild. 2011. Modularity and intuitions in formal semantics: the case of polarity items. *Linguistics and Philosophy*, 34(6):537–570.

Henriëtte de Swart. 1993. *Adverbs of quantification: A generalized quantifier approach*. New York: Garland.

Ewan Dunbar, Gabriel Synnaeve, and Emmanuel Dupoux. 2015. Quantitative methods for comparing featural representations. In *Proceedings of the 18th International Congress of Phonetic Sciences*.

Gilles Fauconnier. 1975. Pragmatic scales and logical structure. *Linguistic Inquiry*, 6(3):353–375.

Dan Garrette, Katrin Erk, and Raymond Mooney. 2014. A formal approach to linking logical form and vector-space lexical semantics. In Harry C. Bunt, Johannes Bos, and Stephen Pulman, editors, *Computing meaning*, pages 27–48. Dordrecht: Springer.

Bart Geurts and Frans van Der Slik. 2005. Monotonicity and processing load. *Journal of Semantics*, 22(1):97–117.

Lila R. Gleitman, Kimberly Cassidy, Rebecca Nappa, Anna Papafragou, and John C. Trueswell. 2005. Hard words. *Language Learning and Development*, 1(1):23–64.

Betty Hart and Todd R. Risley. 1995. *Meaningful differences in the everyday experience of young American children.* Baltimore: P. H. Brookes.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Lisbon, Portugal, September. Association for Computational Linguistics.

Karl Moritz Hermann, Edward Grefenstette, and Phil Blunsom. 2013. "not not bad" is not "bad": A distributional account of negation. *Proceedings of the ACL Workshop on Continuous Vector Space Models and their Compositionality*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Michael N. Jones, Walter Kintsch, and Douglas J.K. Mewhort. 2006. High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, 55(4):534–552.

Saul A. Kripke. 1959. A completeness theorem in modal logic. *The Journal of Symbolic Logic*, 24(1):1–14.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Language Learning*, pages 171–180.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Mike Lewis and Mark Steedman. 2013. Combining distributional and logical semantics. *Transactions of the Association for Computational Linguistics*, 1:179–192.

David Lewis. 1975. Adverbs of quantification. In Edward L. Keenan, editor, *Formal semantics of natural language*, pages 178–188. Cambridge University Press.

K. Lund and C. Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, 28(2):203–208.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *Proceedings of ICLR*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, pages 746–751.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 12:1532–1543.

Martin Redington, Nick Chater, and Steven Finch. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4):425–469.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden, July. Association for Computational Linguistics.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Peter D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.