UNIVERSITÀ DEGLI STUDI DI TRENTO

CIMeC - Center for Mind/Brain Sciences

Research Master's Degree in Cognitive Science

Academic Year 2018-2019

Meaning shifts in distributional models: do we speak the way the world is?

SUPERVISOR: Aurélie Herbelot

STUDENT: E. Kuzmenko

CO-SUPERVISOR: Johan Bos

Contents

Abstract

When we talk about objects that exist in the real world, how differently are they represented in our speech compared to their real properties as we perceive them? Which entities have similar representations in our speech but are completely different when it comes to life? In this project we want to investigate how much the world depicted in the distributional space differs from the real world. It was previously shown that the representations for colors of some entities in our speech are modified according to Gricean maxims (Rawee, 2018). Now we want to investigate these differences not only for the color subspace but on a large scale. In addition, current research in the field suggests that one semantic space can be mapped to another semantic space (Vecchi et al., 2011; Herbelot and Vecchi, 2015). However, it may be the case that two spaces differ too much, so a high quality mapping may not exist. Thus, in our research we also want to investigate how well the linguistic distributional space is mapped to the world space.

In order to perform such a comparison, we build a model representing the real world from the Visual Genome dataset (Krishna et al., 2017), a large database of images marked with objects, attributes, and relations. The language use space is built from the English Wikipedia corpus, and the British National Corpus (BNC). Exploration methods include nearest neighbor comparison and mapping frequency lists between spaces. The comparison of spaces shows that there are significant differences in concept representations between spaces, resulting in semantic shifts for the majority concepts in the language.

1 Introduction

In recent years, distributional semantics models (DSMs) (Erk, 2012; Clark, 2012; Turney and Pantel, 2010) have received close attention from the linguistic community. One reason is that they are known to produce excellent representations of lexical meaning, which account for similarity and polysemy and correlate well with a range of behavioural data (Lenci, 2008; Mandera et al., 2017). Though distributional models have proved to be useful for many NLP tasks, recently there have begun a discussion about the quality and the nature of semantic representations they provide (Westera and Boleda, 2019). One of the problems the discussion raises is the correspondence between the information encoded in the distributional model and the state of things in the real world. Does language use reflect accurately the experience of the speaker with regard to the world they live in? Do behavioral data used for evaluating semantic models reflect this experience? How do these three aspects – experience of the world, language use, and elicited behavioural data – correlate with each other? All these questions remain unanswered in current research on distributional semantic models. Meanwhile, the information encoded in the models is taken as is, without investigating its correspondence to the real world experience.

In their core, distributional models are built over corpus data which may or may not reflect 'what is in the world' (Herbelot, 2013) – and consequently does not reflect human experience gained from the real world data (Andrews et al., 2009). Nevertheless, previous studies have shown that similarities of concepts extracted from distributional models coincide with behavioural data, such as MEN and SimLex-999 datasets (Bruni et al., 2012; Hill et al., 2015). In these datasets, humans were asked to grade how similar are concepts in a pair. When we induce cosine similarities of concepts from a distributional model, we find out that these similarities correlate quite well with at least the MEN dataset (with Spearman ρ equal or higher than 0.7 for different model architectures). This shows us that distributional semantic models do capture the semantic notion of similarity in the way close to conceptual knowledge.

However, if people consider that *cat* and *dog* are very similar according to their perceptual knowledge, and they turn to be very close in a distributional model as well, does it mean that they are indeed very similar in the human perception of the world, i.e. in the very stimuli a speaker is exposed to? When people use natural language, concepts from the real world undergo a perceptual shift as people tend to talk about some properties of concepts and omit other ones. In the case of *cat* and *dog* it may be more important for the speaker that they are both domestic animals and are met in similar situations, and the fact that they belong to two different species is omitted. Thus, concept representation that we find in a corpus may be completely different from stimuli a person perceives. This perceptual shift is averaged across language uses of different people as the corpus data combines texts created by various authors.

For example, in the previous research it was shown that colors of objects extracted from corpora are different from the ones we find in the real world, and also do not correspond to human judgments (Rawee, 2018). Thus, the most salient colour for elephants in the language data was found to be *pink*, whereas in the reality we cannot find pink elephants.

Our present research is therefore aimed at investigating two questions: first, if we try to build a distributional semantic model that approximates concept features as they are in the real world, will it also correspond with behavioral data like corpus-based

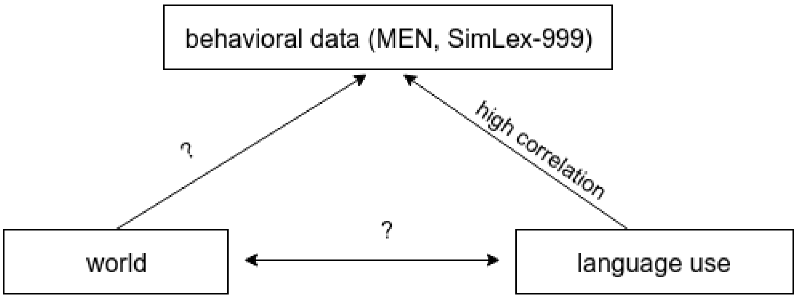


Figure 1: A scheme of the research design.

semantic models? Second, if we do find that similarities between concepts in our world-based space correlate with similarities in the behavioral datasets, does it mean that the world-based space is similar to the corpus-based space? Do the concepts in both kinds of spaces possess similar representations? Does this similarity pertain to spaces as a whole or to particular representations within a space?

These questions are important because previous literature has assumed the existence of a learnable function mapping between the space of language use and the space of real world features. However, the complexity of such a function has been little discussed (one of few studies discussing it is (Rubinstein et al., 2015)). Overall, the present research does assume that different semantic spaces can be mapped to each other – in other words, you can apply some transformations to one space and arrive to another space (Vecchi et al., 2011; Lazaridou et al., 2014; Herbelot and Vecchi, 2015). Thus, the hypothesis is that if two spaces possess similar representations of the same concepts, they are similar as a whole and can be mapped. On the other hand, if some concepts in two spaces turn out to be similar, it does not mean that it applies to the whole space. In this case the hypothesis about similarity of spaces does not hold. If we find out that this applies to the created world-based and corpus-based spaces, we also aim at investigating what kind of semantic shifts concept representations undergo in the corpus-based models compared to the real world. Overall, our hypotheses are described on Figure 1.

We consider this problem important from many points of view:

- finding out which categories of concepts undergo meaning shifts and which categories stay the same may help to understand how semantic information is stored in the human conceptual space;

- finding out what properties of concepts change in the language use may help to understand what people consider important for communication;

- finding out which components of meaning are more discriminative for the “model of the real world” may help us look critically at the output of existing semantic systems.

Several possible reasons why our language space differs from the real world space can be:

- we don’t talk about the world as it is because we describe all things in it through the prism of our perception. This, human tend to annotate in the VG data and talk in the language about things that are salient and important for them. For example, in the VG data eyes of people or animals are frequently annotated, whereas ears are hard to find in the annotation.

- we don't talk about the world as it is because we're Gricean and thus do not want to talk about obvious and unimportant things (Grice et al., 1975);

- we don't talk about the world as it is because of human creativity, as we create other possible worlds with our language.

Thus, the objective of our work lies in the comparison of the representations of concepts in the real world and in the language model. To perform such a comparison, we need a model that reflects state of things in the real world. As it may seem impossible to model object properties in the real world directly (describe colors, and texture, and size, and actions of every concept), we can in fact use truth-theoretic models of semantics as a proxy to the real world. In truth-theoretic models meaning is defined as a correspondence relation between predicates and the world. They are also built from entities and sets rather than from *word* co-occurrences, that is, from lexical types like corpus-based DSMs. Therefore, formal models account for denotation and set-theoretic aspects of language, but they are often said to lack the ability to account for lexical similarity and gradedness. This has been the basis for wanting to combine formal and distributional semantics in the past (Boleda and Herbelot, 2016): the role of DSMs, it is claimed, is to bring the lexicon to denotational approaches to meaning. In the present paper we create a truth-theoretic model that could also account for distributional properties of entities and use this model as the proxy to the real world. The process of building such a space and evaluating it is described in Section 4.

By comparing the distribution of concepts in our truth-theoretic data with the distribution of corresponding word occurrences in the language data we are hoping to find out the underlying semantic differences between the two spaces. The comparison of a world-based space to the corpus-based spaces is described in Section 5.

The research goals we are trying to achieve are therefore as follows:

- create a model of the real world based on the combination of the truth-theoretic and distributional data;

- explore the properties of this model and the differences between this model and corpus-based models;

- investigate the similarity of language space and the world space and show that representations of concepts in both spaces differ significantly;

- find out what properties of concepts are represented similarly in the language space and the world space;

- create a map of different categories of concepts and show on the map which categories shift more than others;

- describe in details the nature and direction of semantic shifts found in the corpus space;

- figure out what transformations representations of concepts undergo when moving from real world space to language-based space.

The overall conclusions of the research can be described in the following way: we can find certain differences between concept representations in the semantic space built from the real world and in the semantic space based on corpora (i.e. natural language). In other words, the meaning of concepts we use (reflected in their distribution in the corpora) differs from the composition of their properties in the real world. This may be true not only for the color subspace (Rawee, 2018), but for other dimensions as well. Some particular categories of concepts may differ more significantly in both spaces than other categories. Thus, some concept representations are more stable in the corpus space as compared to the real world-space whereas other undergo a significant semantic shift.

The structure of the thesis is as follows: first we describe the process of building a truth-theoretic space which approximates the real world. After building such a space, we study its properties and describe the contributions of different features into the overall quality of the model. Relating to Figure 1, these steps reflect the left arrow on the diagram. After that we have a deeper look into the particular differences between corpus-based and world-based models. This direction of research corresponds to the arrow at the bottom of Figure 1. First we investigate differences in the vocabulary of the models – looking for things that are absent in the real-world model and present in the language. Then we have a look at the differences in the salience of objects. The next step is to compare neighbor lists for concepts in both corpora, which may help in describing the nature of semantic shifts they undergo.

2 Related Work

For the purpose of this work, I will give an overview of current research and methods in three areas – distributional semantics, formal semantics and space mapping. Distributional semantics is the main tool I am using for my research, and when trying to establish the correspondence between the behavioural data and the world, I also apply some methods of formal semantics. The review of the research on space mapping helps to understand how we can find a link between world and language, and in what sense direct mapping between these two spaces can be problematic.

2.1 Distributional semantics

In natural language processing, distributional models, based on the foundational idea of ‘meaning as context’, are now one of the primary tools for semantic-related tasks. They stem from the so called **distributional hypothesis**, which states that co-occurrence statistics (word co-occurrence distributions) extracted from large enough natural language corpus can in some way represent the ‘meaning’ of words as perceived by humans Firth (1957). More formally, with a given *training corpus*, each word is represented as a vector of frequencies for this word occurring together with other linguistic entities (its contexts). These vectors are located in a *semantic space* with all possible contexts or semantic features as dimensions. Vector models of distributional semantics or vector space models (VSMs) are well established in the field of computational linguistics and have been studied for decades; see the review in Turney et al. (2010), among others.

Recently, a particular type of these models has become very popular, namely, *predictive models* introduced in Bengio et al. (2003) and Mikolov et al. (2013). They are implemented in the wide-spread *word2vec* software tool. Predictive models directly learn dense vectors (*embeddings*) that maximize the similarity between contextual word pairs found in the data while minimizing the similarity for unseen contexts. However, in my study I concentrate on the more traditional count-based models and use predictive models only occasionally.

Distributional models compute meaning by analysing word co-occurrences. The trained model can represent semantics of a given word as a sequence of its 'nearest neighbours': words closest to the key word by the cosine similarity of their vectors (embeddings).

Though distributional models are widespread and were previously studied in many details, the distributional models created for our project from the Visual Genome dataset are novel with respect to the data we use. However, some research already suggested using multimodal and image datasets to model semantics. Most similar to our endeavour is the work by Young et al. (2014), who also take multimodal datasets as a basis to learn denotations. Their model is however created for the task of semantic inference and takes the extension of a word to be the set of situations it applies to. We introduce notions of entities and properties in our own model. Another research about semantic entailment from multimodal datasets is (Hürlimann and Bos, 2016), where the authors use semantic representations built from images to predict spatial relations between objects.

We also overlap with work on DSMs using structured and/or encyclopedic resources (Faruqui and Dyer, 2015; Recski et al., 2016), which adds a 'real world' dimension to distributional models.

2.2 Formal Semantics

Our work fits into attempts to bridge the gap between distributional and formal semantics. The subfield of Formal Distributional Semantics (FDS, Boleda and Herbelot, 2016) includes efforts to a) investigate the mapping from distributional models to formal semantic models (Herbelot and Vecchi, 2015; Erk, 2016a; Wang et al., 2017a); b) enrich formal semantics with distributional data (Garrette et al., 2011; Beltagy et al., 2013); and c) account for particular logical phenomena in vector spaces, including composition (Coecke et al., 2011; Boleda et al., 2013; Baroni et al., 2012; Bernardi et al., 2013; Asher et al., 2016 amongst many others). We also note the relevance of the work on constructing distributional spaces from syntactically or semantically parsed data (e.g. Padó and Lapata, 2007; Grefenstette et al., 2014; Hermann and Blunsom, 2013), which echoes the way we construct vector spaces from various types of predicative contexts. In contrast to those efforts, however, our data is not a standard corpus reflecting word usage but a collection of logical forms expressing true sentences with respect to a model of the world. We also note the related work by Larsson, 2013, who proposes a formal framework for defining meaning from world perception.

2.3 Space Mapping

As I have pointed out in the introduction, the current research in the field suggests that two semantic spaces can be directly mapped to each other. Examples of such straightforward mappings can be found in various papers. In (Fagarasan et al., 2015) the authors try to ground distributional semantic models using visual data. In particular, they create a linear mapping between a distributional space and a space build from feature norms. Using this mapping, they further predict feature norms for unknown concepts. However, the authors do not analyze the performance for different classes of concepts, therefore, it remains unclear how well the feature norms are generalized from the learned mapping.

In the paper by (Gupta et al., 2015) the authors try to predict referential attributes of a concept from its distributional representation. The authors learn a straightforward linear regression for predicting the features, and it also remains unclear in which situations this methodology works well and in which it doesn't work at all. The authors themselves suggest that some groups of attributes are predicted better than other ones (*geolocation* and *population* are the best attribute groups) and that many features are hard to predict as they are culturally biased.

In (Herbelot and Vecchi, 2015) the authors also perform a straightforward mapping between a distributional semantic space and a set-theoretic model. The authors try to prove that there is a relationship between distributional information and concept co-occurrences with natural language quantifiers and we can thus predict the quantifier with which the object occurs from a distributional representation. This problem is quite narrowed because quantifiers are taken into account from the wide range of concept features. The authors notice some strange artefacts of space mapping such as the fact that the vector for *alligator* being close to *crocodile* in the mapped space, but not to *alligator* itself. Overall, this research stays in the frame of the possibility of a direct mapping from one space to another.

Recently there appeared a body of research suggesting that the relationships between two spaces, in particular between a language space and a space of concept features are more sophisticated. In (Erk, 2016b) the authors try to find out the kinds of semantic information between that a model can distinguish. They also argue that though a direct mapping is not plausible, we still can approximate concept features in one space (e.g. world space) by learning its features in another space (e.g. language space). In other experiments authors have applied Bayesian methods to learning features of concepts from a space (Wang et al., 2017b).

A start of a criticism for direct mapping between spaces in general and between information encoded in distributional models and world experience is described in (Rubinstein et al., 2015). Here the authors argue that distributional properties do not encode equally different types of semantic information. Thus, it is not possible to successfully predict attributes for various concepts from their distributional representations.

3 Data

We perform our analysis on the basis of the three spaces. One of the spaces is built from the data in the Visual Genome and represents the world space, and the other two are traditional corpus-based models created from the British National Corpus and the English Wikipedia corpus representing the language use.

The Visual Genome dataset (Krishna et al., 2017) is used to build a large set-theoretic model as an approximation of “the real world”, and show that quality vector representations can be extracted from it and evaluated in a way similar to traditional corpus-based spaces.

To obtain our model, we take the annotation of the Visual Genome (henceforth VG),

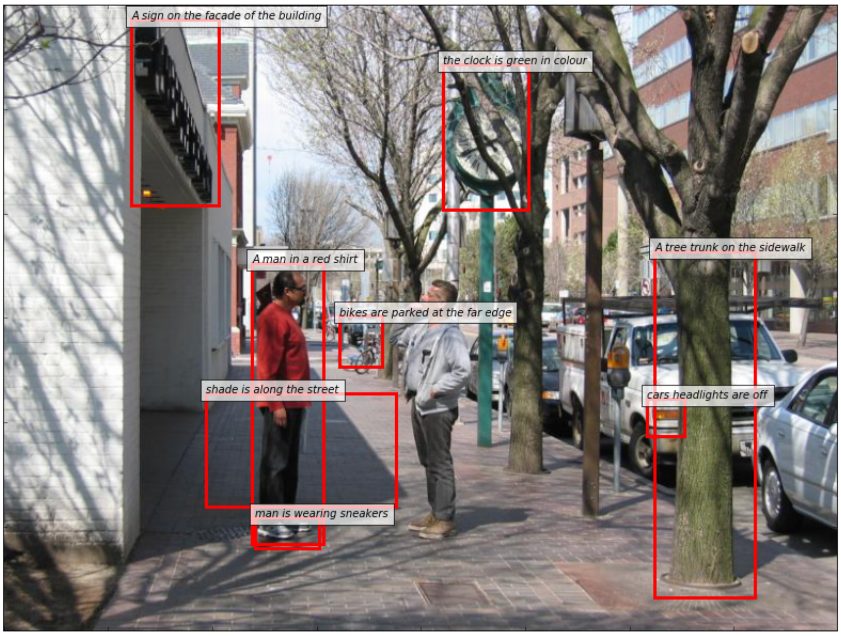

A sign on the facade of the building
the clock is green in colour
A tree trunk on the sidewalk
A man in a red shirt
bikes are parked at the far edge
shade is along the street
cars headlights are off
man is wearing sneakers

Figure 2: An example picture from the Visual Genome dataset. Object frames are colored in red, object types are specified.

a large database of images annotated with objects, attributes and relations, and regard this data as an informative, although incomplete, description of the world. The VG dataset consists of 108077 images annotated with 3.8 million objects present in the images and 2.8 millions of their attributes. Another important type of annotation present in the VG is relationships between objects, and there are 2.3 million relationships annotated in the dataset. An example of an annotated image is shown on Figure 2.

We convert the annotated data into a representation akin to some underspecified first-order logic. From this representation, we build several DSMs from various aspects of the representation and inspect the properties of the created spaces. This process is described in Section 4.2. In the created space every entity is characterized by the following dimensions: its attributes, relationships it is engaged in. The previous research has shown that the spaces built from different subtypes of the VG data have various quality (Kuzmenko and Herbelot, 2019).

However, the annotated objects and their properties are subject to the perception of human annotators that created the dataset. The annotations themselves represent linguistic utterances and are therefore represent the real world already through our linguistic prism. However, we still consider the model built on these data close to the state of things in the real world.

The corpus-based models are built from the BNC data and English Wikipedia. The BNC corpus consists of 85 million words and includes different text genres and styles and thus gives us a more “general” impression of a language space. English Wikipedia corpus consists 40 million words and is supposed to be more encyclopedic-like and thus are more likely to encode real distributional properties of concepts. Wikipedia corpus was created from scraped Wikipedia pages in November 2018. Though it is possible to find a newer version of English Wikipedia, we suppose that language phenomena did not change significantly since then and the present corpus is still appealing to our needs.

We suppose that Wikipedia space and the VG space will be more similar to each other than the BNC space and the VG space.

We process corpora in the following way: the texts are split into sentences, the sentences are tokenized, taking punctuation signs into account. After that we build count-based distributional spaces with our own script, with no lemmatization or any additional pre-processing. The script counts word co-occurrences in a sentence withing a window of size 3. That is, for every word we take its immediate neighbors from left and right. Also, we create representations for 10 0000 most frequent words as we are not interested in words with low frequency. Thus, the dimensionality of the BNC and Wikipedia spaces is 10 000. Obviously we meet the problem of different vocabularies in the VG space and corpus-based spaces since corpora contain much more different lexical types. To cope with this problem, we additionally created semantic spaces containing only those types and dimensions that are present in the VG vocabulary. That is, if there is an attribute 'purple' in the VG, but no attribute 'turquoise', we drop the dimension 'turquoise' from our corpus-based spaces. Additionally, all the concepts that are not met in the VG are also dropped. Most of them are abstract concepts such as 'happiness' or 'life', but also some concrete objects that are not met in the VG (e.g. 'apricot'). The resulting data are stored as sparse matrices, from which we later extract all the needed data.

4 Building the Real World Space

4.1 Introduction

4.2 Experimental Setup

This part of our work is concerned with two problems: first, is it possible to build a large set-theoretic model as an approximation of "the real world"? Second, how does such a model compare to behavioral data (i.e. traditional evaluation datasets) an corpus-based spaces? Therefore, in this chapter we are concerned with the left edge of the semantic triangle described on Figure 1.

In order to build a "real world" space, we require a representation akin to a set-theoretic model. We take the annotation of the Visual Genome (VG) dataset (Krishna et al., 2017) as a proxy for such model, under the assumption that it provides a set of 'true' sentences about the world. There are three types of annotation in the VG dataset: a) entities, or **objects** (e.g. *'window'*, *'elephant'*) – the individuals present in a given image; b) **attributes** (e.g. *'red'*, *'made of bricks'*) which describe the properties of objects; c) **relationships** (e.g. *'on'*, *'has'*, *'wearing'*) which correspond to relations between objects. The dataset also includes situations, or **scene graphs**, which correspond to a single image and describe all the objects that co-occur in that image. Thus, on Figure 2 a situation on the image contains a tree, a car, a man, a sidewalk and a shade (from the tree), associated with bounding boxes. We do not use the image itself but solely the annotation data from the dataset graph.

It should be noted that every object in VG is assigned a WordNet synset and a unique id. This allows us to pre-process the data and convert it into shallow logical forms corresponding to predicate / entity pairs, ordered by situation and implicitly coordinated by a $\wedge$ within that situation. For instance, the following toy example indicates that situation 1 contains a tall brick building, identified by variable 1058505 in VG, on which we find a black sign, identified by variable 1058507. Note that the identifiers are 'real-world' variables, which pick out particular objects in the world.

$$\begin{array}{ll} S1 & building.n.01(1058508), tall(1058508), brick(1058508) \\ & sign.n.02(1058507), black(1058507), on(1058507{,}1058508) \end{array}$$

This representation allows us to capture all the distinct objects annotated with e.g. the synset *'building.n.01'*, and then we can generate the set of buildings (*building′*) in our universe.[1] To avoid data sparsity, we convert all relations into one-place predicates, by replacing each argument in turn with its corresponding synset. So in the example above, $on(1058507{,}1058508)$ becomes $on(1058507{,}building.n.01)$, $on(sign.n.02{,}1058508)$, which formalises that 1058507 is in the set of things that are on buildings, while 1058508 is in the set of things that signs are on.

Formally, the VG data can then be considered a set-theoretic model $M =< U{,}I >$ where U is the universe (the set of all objects in the model, as identified by 'real-world' variables), and I is an interpretation function mapping from a set of n-place predicates P to n-tuples of objects in U (with $n = 1$ given our pre-processing of relations). P is the union of synsets (Syn), attributes (Att) and relations (Rel) in VG. We then build a distributional space $S =< U{,}P{,}D{,}F{,}A{,}C >$ where U and P are the universe and the predicates as above; D are the dimensions of the space so that $D \subseteq P$ (that is, any combination of Syn, Att and Rel); and F some extraction function over our corpus of shallow logical forms C. F is of the form $U \times D \rightarrow \{0{,}1\}$, i.e. it returns whether a particular dimension is predicated of an entity, giving us boolean entity vectors for all objects in VG. Finally, an aggregation function $A : (U \times D \rightarrow \{0{,}1\}) \rightarrow (P \times D \rightarrow \mathbb{N}_0)$ returns the final space by summing the entity vectors corresponding to each predicate in P: $\mathbb{N}_0$ is a natural number expressing how many times dimension D is predicated of entities of type P. The summing operation follows the model-theoretic intuition that a predicate p denotes a set which is the *union* of all things that are p: for instance, all dog entity vectors are summed to produce a vector for the predicate dog′.

In addition, we consider two ways to augment this original setup. The first way is to add situational information to the representations: while relations tell us what type of things a particular entity associates with via a particular predicate, this information does not include the type of things the entity simply *co-occurs* with. For instance, we may have a situation where a dog interacts with a ball (encoded by some relation *dog - chew - ball*), but VG relations do not directly tell us that the dog entity co-occurs with a *park* entity or a *cloud* entity. We can add this situational information simply by counting the number of times each entity co-occurs with entities of other types, without a relation being explicitly expressed. This approach also allows us to account for the "big picture" and not only for the immediate features of a given object, which is more similar to working with sentences from a natural corpus.

Another way to augment the data is to add encyclopedic information to the VG data, which could be part of a more 'complete' model including some generalizations over the encoded sets. To do this, we extract hypernyms from WordNet (Miller et al., 1990) using the *nltk* package.[2] Only one level – the immediate parents of the concept – is taken into account. We note that hypernyms are different from the other VG features in

[1] Note that we are not making use of the sense information provided by the synset in this work. Most words in VG are anyway used in a unique sense.

[2] `http://www.nltk.org/`

that they don't come from natural utterances (no one would say *"domestic animal"* in place of *"dog"* in a natural context).

Having prepared the data, we build variations of the model M by counting co-occurrences between our entity set (aggregated into predicates with function A) and the following features $D \subseteq P$: attributes (*Att*), relations (*Rel*), situations (*Sit*), hypernyms (*Hyp*), as well as all their combinations.[3]

4.3 Evaluation of the models

Having built the world-based models, we are now concerned with their evaluation, since it both tells us how do the representations agree with human perception expressed in behavioral datasets, and how do these models compare to traditional corpus-based models.

We evaluate our models with both relatedness and similarity datasets (MEN, Bruni et al., 2012, and SimLex-999, Hill et al., 2015), drawing conclusions about the actual behaviour of world-based models with respect to behavioural data. The MEN dataset is supposed to capture the relatedness notion, which is defined as the relation between pairs of entities that are associated but not actually similar. SimLex-999 accounts for similarity, which is defined as the relation between words which share physical or functional features, as well as categorical information (Hill et al., 2015). Both datasets are structured in the same way: they consist of word pairs human-coded for their level of association, which is the average value produced by human annotators when constructing the dataset. They respectively include 3000 (MEN) and 999 (SimLex) word pairs. To evaluate our DSMs, we follow standard practice and compute the Spearman ρ correlation between the cosine similarity scores given by the model and the gold annotation. Results are shown in Table 1. To maximise comparability between different spaces and with text corpora, scores are given for raw co-occurrence matrices, and no dimensionality reduction or other optimization of the space is conducted. Note that due to the size of VG, we cannot evaluate on all pairs in the datasets. We show actual coverage in brackets next to the correlation scores.

Trends are similar both for MEN and SimLex-999. We get overall best results (highlighted in **bold**) for the models built using relations, situational information, and relations together with situations. Other models have significantly lower quality, both for single features and for their combinations. It should be noted that taking all the features together does not improve the quality of the space, as it seems that there are counterbalancing forces from relations and situations improving the score and other features taking the score down.

In the last column of Table 1 we report the total number of co-occurrences in each variation of the world-based model. They are included in order to make sure that we do not observe solely the effect of increasing the amount of data. Indeed, models with the greatest number of co-occurrences show medium quality, and for some combinations of features the score even decreases with more data (e.g., compare the *Hyp* and *Hyp + Sit* models, where the MEN score stays more or less the same and the SimLex score becomes lower). Moreover, the *Rel* model shows the highest score on a moderately small amount of data for the MEN dataset, and for the SimLex-999 dataset the score is a bit lower, whereas the *Rel + Hyp* model becomes the best (though hypernyms come from outside the model).

[3] The code to pre-process the Visual Genome and the data to reproduce the experiments can be found at `https://github.com/lizaku/dsm-from-vg`.

Setting	**MEN**	**SimLex-999**	**Num. co-occurrences**
Attributes (*Att*)	0.1801 (871)	0.1119 (217)	1 854 033
Relations (*Rel*)	**0.5499** (847)	**0.2861** (216)	6 481 872
Situations (*Sit*)	**0.5294** (847)	**0.2480** (216)	22 894 730
Hypernyms (*Hyp*)	0.3399 (956)	0.2128 (244)	1 989 576
Att + *Rel*	0.346 (871)	0.1840 (217)	10 720 260
Att + *Sit*	0.4492 (871)	0.2042 (217)	25 988 265
Rel + *Sit*	**0.5326** (847)	**0.2463** (216)	32 170 563
Att + *Hyp*	0.2385 (975)	0.2055 (244)	5 114 997
Rel + *Hyp*	**0.5193** (956)	**0.2979** (244)	10 878 274
Hyp + *Sit*	0.3860 (956)	0.1731 (244)	26 882 218
Att + *Rel* + *Hyp*	0.3430 (975)	0.2367 (244)	16 391 743
Att + *Rel* + *Sit*	0.4503 (871)	0.2018 (217)	37 652 176
Att + *Sit* + *Hyp*	0.3260 (975)	0.1319 (244)	31 252 206
Rel + *Hyp* + *Sit*	0.3900 (956)	0.1760 (244)	38 571 325
Att + *Rel* + *Hyp* + *Sit*	0.3283 (975)	0.1337 (244)	45 329 361

Table 1: Spearman ρ correlation for various models on MEN and SimLex-999.

To compare performance of our truth-theoretic models with traditional DSMs built from text corpora, we create count-based models from the English Wikipedia using a window of ± 2 words around a target. We modulate corpus size to roughly match the number of co-occurrences extracted from VG.[4] Additionally, we train predictive models with Word2Vec (Mikolov et al., 2013) with the same number of co-occurrences as in the count-based variants. We use the same window size of 2, and the dimensionality of vectors is set to 300. The evaluation scores for different corpora sizes are shown in Table 2. We can see that, in contrast with the VG models, the score for count-based models is dependent on the amount of data provided to the DSM, and generally lower for similar numbers of co-occurrences (scores are consistent with results reported by Sahlgren and Lenci, 2016). Predictive models are simply not able to construct high-quality word representations from such amount of data. The scores of the VG models do not improve with more data, which confirms that the evaluation score of the VG models is not dependent on the number of co-occurrences.

When we try to improve the quality of our best world-based model (*Rel*) by applying normalisation, dimensionality reduction (to 300 dimensions) and PPMI weighting, we reach scores of **0.6539** on MEN (847 pairs are evaluated because not all of the pairs in the evaluation dataset are present in the VG space) and **0.3353** on SimLex-999 (216 pairs evaluated). Whilst results are not directly comparable, we nevertheless note that the MEN score is close to the figure of 0.68 reported for the inter-annotator correlation on the full 3000 pairs.[5] It is also only a few points lower than the best score of 0.72 obtained by Baroni et al. (2014) over 2.6B words (around 1600 times more data than in *Rel* on the basis of a ±2 word window size). The SimLex figure is also well above

[4] Models are built using `https://github.com/akb89/entropix`.

[5] See `https://staff.fnwi.uva.nl/e.bruni/MEN`.

Count-based		Predictive (word2vec)		Co-occurrences
MEN	SimLex-999	MEN	SimLex-999	
0.081 (749)	0.050 (462)	0.024 (749)	0.003 (462)	2 000 000
0.158 (995)	0.010 (546)	0.043 (995)	0.019 (546)	5 000 000
0.225 (1226)	0.038 (610)	0.049 (1226)	0.020 (610)	15 000 000
0.226 (1455)	0.037 (688)	0.031 (1455)	0.046 (688)	30 000 000
0.253 (1554)	0.056 (696)	0.031 (1554)	0.044 (696)	40 000 000

Table 2: Spearman correlation on MEN and SimLex-999 datasets (Wikipedia spaces)

the figure of 0.233 reported by Hill et al. (2015) on an SVD model trained over 150M words ($\approx$ 100 times more data).

4.4 Discussion

In this section I have described the attempts to build a world-based semantic space and evaluate it in a way similar to evaluating standard corpus-based spaces. As we can see from the evaluation scores, our created model performs quite will on MEN and SimLex-999 datasets, thus showing that the conceptual representations elicited from behavioral data are correlated with raw exposure to the world. The highest correlation score ($\rho = 0.5499$) is lower than reported scores for corpus-based spaces ($\rho = 0.72$, (Baroni and Lenci, 2010)), but still shows that our space is comparable to corpus-based spaces. Moreover, we retrieve better semantic representations from a smaller amount of data, as we are not able to achieve high correlation scores on small amounts of corpus data.

Some interesting observations can be made with regard to the type of properties that seem to be relevant to modeling conceptual association. First, the relative results we are observing across the VG models are not artefactual of model size. Thus, a model based on situations, with 22M co-occurrences, performs worse than the model with relations, which comprises only 6M co-occurrences. This tells us that some aspects of the model-theoretic data are much more important than others and that some can even be detrimental. This finding echoes results in Emerson and Copestake (2016), which indicated that selecting particular relations from parsed data can improve performance on SimLex.

Second, the VG models outperform the standard spaces by a large margin on SimLex, even with small amounts of data. This confirms that SimLex encodes a notion of similarity that is better captured by looking at how things 'are' truth-theoretically rather than what we say about them. The fact that attributes perform badly on that dataset, however, contradicts the idea that SimLex encodes similarity of intrinsic features. Indeed, *relations* outperform any other combination of features, showing that how things associate with other things may be more important than how they intrinsically are.

Third, an additional point can be made about relations and situations. While both *Rel* and *Sit* models perform well on their own, the combined *Rel* + *Sit* model has lower quality (around two points are lost on MEN and four points on SimLex, compared to *Rel* alone), which means that situations take the score down. This can be explained by the fact that situations are a "noisy superset" of relations: some of the entities that

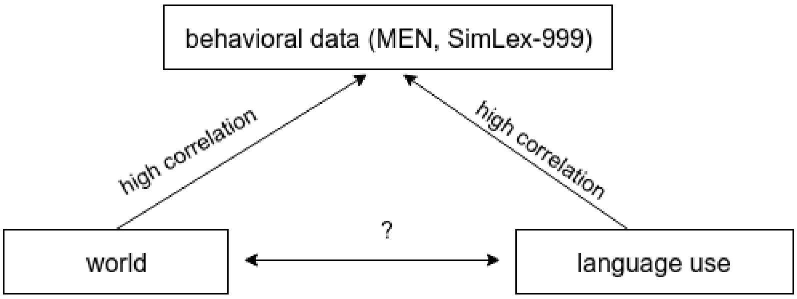


Figure 3: A scheme of the research conclusions.

co-occur in a situation will have an explicit relation associated with them (e.g., *cat* and *mouse* related by *chase'(x,y)*), while others may indeed solely co-occur (e.g., *cat* and *fork* in a scene with a pet sitting next to a dining table). So it seems that aspects of the world that entities are *actively* involved in are more important to define them than simple 'bystander' individuals.

Finally, using hypernyms improves the quality of models when evaluated on SimLex. This confirms previous results showing that using dictionaries and lexical databases helps getting better performance on SimLex (Faruqui and Dyer, 2015; Recski et al., 2016). It also indicates than when computing similarity, humans may indeed activate some 'meta-knowledge' which is not directly encoded in the basic level categories (Rosch et al., 1976) people use to describe a situation.

Overall, we have shown that the created space features high-quality semantic representations, and the differences between various tested model architectures point us to the fact that different components of meaning can be captured in distributional properties of concepts. However, in general there is high correlation between the world model built on the VG data and the behavioural data depicted in MEN and SimLex-999 datasets. Thus, the first part of our initial research hypothesis is verified, and the current state of our results is reflected on Figure3. However, it is yet unclear what it means for the correspondence between world exposure and language us. Therefore, in the next section I am going to investigate the second part of our hypothesis. In particular, we will see how the captured semantic properties differ between the world-based space and corpus-based spaces.

5 Language Spaces and Representation Shifts

5.1 Experimental setup

In this chapter I am going to have a deeper look into similarities and discrepancies between the real-world distributional space and corpus-based distributional spaces. Thus, we are going to have a deeper look at the lower edge of the semantic triangle depicted on Figure 1.

As we have said in the introduction, corpus-based and world-based models can differ fundamentally in the way they express semantics. Whereas the real-world space is built on the basis of denotation (i.e. we have specific entities and their types as annotated by people), the corpus-based space operates on lexical occurrences, which are not bound to specific objects. In other words, when naming concepts in the VG annotation, people decide on the name based on their perception of a particular concept. Thus, an object X can be called ‘a dog’ if people believe it is a dog, and it can have attribute ‘black’ if it is perceived as black. At the same time, the corpus-based space reflects the same objects but through an additional prism of language. ‘A black dog’ can turn to a more specific instance of a shepherd dog, because people decided to make it more concrete. A corpus-based space also features metaphorical uses, and the phrase *be in the doghouse* may mean different things in the world-based space (*physically be in the doghouse*) and in the language-based space (*be in disgrace*).

To make the comparison of the spaces more fine-grained, we build language models from two corpora – BNC and English Wikipedia. The process of building the models is described in Section 3. However, for some experiments we do not employ the distributional spaces at all. Interesting conclusion can be made on the basis of the vocabulary of both worlds and concept salience (i.e. what concepts are met in VG, what concepts are met in corpora, which concepts are more salient in VG, but not that important in the corpora). For these experiments I use frequency lists of VG, BNC, and Wikipedia

When analyzing the spaces, we perform the analysis in three directions – thus, I conduct three experiments. The first experiment aims to investigate the differences in the vocabulary between the spaces. As it is natural that different spaces contain representations for different words, it could be useful to know which words are not present in the VG space. It can give us a clue about things that are existing in the language but are not used to describe the real world (in accordance with, for example, the prototype theory (Rosch et al., 1976; Rosch, 1999)). To find out the differences in the vocabulary, we need to compare the dimensions that do not overlap between spaces. Some obvious things that are not in the VG space are abstract things, but we aim to concentrate on concrete nouns that could be present in the VG data, but are in fact absent. The main hypothesis that we want to check when comparing the vocabularies is that in the language space objects that we see every day should be less frequent than in the VG since people are trying not to talk about obvious things (Grice et al., 1975).

It should be noted that we are not looking for the limits of the Visual Genome vocabulary. It is expected that the number of objects and their features annotated in the dataset is limited. Thus, any word absent in the VG is bound to be found in a bigger corpus, such as BNC or Wikipedia. However, we aim to investigate which things hypothetically should fall within those limits but are still missing. This will point us to the conceptual differences between the spaces.

The second experiment includes analyzing overlapping dimensions for words, which should reflect differences in distribution. Whereas the previous analysis concentrates on the things that are described in the language model but not in the world model, this approach looks into concepts that are represented in both spaces but in different ways.

In this second experiment we are comparing frequencies of concepts in the VG space and corpus-based spaces. The hypothesis we are checking in this case is that if an object is frequent in the real world, it should appear with high frequency in the natural language as well. Vice versa, the objects that are not frequent in the real world (therefore, are not frequently met in the VG annotation) should be less frequent in the corpora as well. Thus, frequencies of objects in different spaces should correlate. Of course, this correlation is not expected to be ideal. We also expect that words from some semantic areas would produce better correlation than other semantic areas. Thus, the correlation scores will be heterogeneous for different semantic clusters.

The third experiment includes computing neighbors for overlapping dimensions to check for specific subspace transformations across models. If some concepts are present both in the world model and the language models, but their distributional properties are still entirely different (even despite similar frequencies), we expect that this will be reflected in their nearest neighbors lists. To extract the nearest neighbors of a word in a distributional model, we calculate cosine similarity between the word in question and other words in the vocabulary and rank the words by this score. In our experiments we extracted 20 most similar words for each dimension. This number was chosen empirically, as 10 most similar words seem to be not enough for a thorough analysis, whereas bigger amounts of neighbors tend to produce a lot of redundant information.

5.2 Experiment 1: Comparing Vocabularies

To find out the differences in the vocabulary, we perform the comparison of non-overlapping dimensions between the spaces. The workflow is as follows: first we compute the list of non-overlapping dimensions in both VG space and the corpus-based spaces (BNC and Wikipedia). After that, we filter out the concrete nouns as concrete objects are supposedly likely to be present in the truth-theoretic model. We also use a frequency threshold to filter out the least frequent nouns. In the end we end up with a list of 500 nouns that are frequent in the BNC and 500 nouns that are frequent in Wikipedia. These should be the entities that are present in our speech but are not mentioned by the annotators of the VG data.

We filter out the concrete nouns on the basis of the list of abstract and concrete nouns created by P. Turney for metaphorical sense identification (Turney et al., 2011). This list was produced by training a classifier on a subset of nouns annotated with their concreteness score, and later this classifier was applied to a vocabulary of 114501 words. Each word is rated from 0 (highly concrete) to 1 (highly abstract). From this list we extracted nouns with the score equal or lower than 0.45. This threshold was determined empirically through experiments. Lower scores were reducing the list of possible nouns, and higher scores resulted in a heterogeneous list of concrete and abstract nouns. The full list of non-overlapping dimensions selected for further analysis can be found in Appendix A.

First of all, we can perform a manual analysis of this list. Among the things that could be present in the VG we can find mainly human roles (*father, mother, aunt*) or professions (*clerk, nurse, doctor, poet*). We can hypothesize that in the language space the speakers are using more specific referring expressions when describing people, whereas in the real world they use more generic terms. Apart from this category of nouns, the corpus-based space also features institutions and area formations that could also be naturally met in the VG dataset (*university, library, village*), however, they are lacking. This could mean that people are also more specific when describing the institutions in the natural language. It may seem justified, as in the real world it is hard to predict the role of a building from its visual appearance (attributes in the VG), and the same idea can be applied to people.

Overall, we can tell that these discrepancies in naming concepts may happen because you cannot tell the difference just by the visual appearance of an object and decide to use a more general referring expression in the real world situation. Another explanation is that people are less committal when describing a picture in a real world situation. Thus, when we spot various flowers or birds, we name them with general terms such as simply “flower” or “bird”. However, when it comes to more specific situations in the language use, we can name particular species of flowers or birds. This explains why in the VG we don't find many occurrences of "sparrow" or "pigeon" though they are apparently frequent in the situations captured on the images. These findings also echo the prototype theory, which suggests that some concepts are less likely to be used when naming an object. This will be investigated in more detail in the upcoming Section 5.3.

When speaking about non-overlapping dimensions between the VG and Wikipedia, we find many animal (*sparrow, tiger, zebra*) and plant (*dandelion, oak*) species in Wikipedia space that are absent in the VG. This trend is similar to what we have seen when looking at the lists for the BNC. Therefore, the main conclusion of this experiment is that in the natural language people differentiate between different kinds of similar concepts (human roles, plant species, types of buildings) that do not differ with regard to their attributes in the real world and thus are annotated with more generic types in the VG (*person, flower, bird*). It may signify, however, that annotators of the VG are just cautious to annotate specific types for the entities, but this still reflects the idea that in the world space the concepts differ less. Another conclusion of this analysis is that in general the natural language tends to have richer vocabulary, even if we take a portion of texts comparable in size with the annotation of the Visual Genome.

Overall, the results of this experiment tell us that one of the reasons for semantic shifts between corpus-based and world-based models is that the concepts are influences by human perception in different ways. In a real world situation we consider some things less important, for example, naming a particular species of a flower is not important at all. In a language situation this information may suddenly become important, thus, the language space features more concrete naming of different phenomena, whereas the world space gives us only general types. With regard to the hypotheses stated in the beginning of this work, I think that this experiment is more likely to prove that the language space differs from the world space because of perceptual differences in various communicative situations and modalities.

5.3 Experiment 2: Comparing Frequencies

Another method that we use for the comparison of the vocabularies is trying to predict which words should be salient in the VG data among the words that we filtered out during for the first experiment: concepts present in the corpus, but absent in the VG data annotation. Salience in our experiments is defined as "importance of a concept in the overall picture of the world". For example, we can hypothesize that people are more salient (i.e more important) in language as well as the real world, thus, they are met in the word occurrences or annotated in images quite frequently. We measure salience as frequency of a word in the corpus or an entity in the VG. Therefore, we are trying to find out not only which words should be present in the VG, but are not there, but also we want to point out the words which should be important (and frequent) in the VG, but are not met at all.

To do this, we compute frequency for overlapping words in the dimensions of the VG space and the corpus-based space (here we experiment with the BNC). Both frequencies are normalized by size of the data as we compute occurrence per million for every word. Having done that, we build a regression that predicts frequency of a word type in VG given its frequency in a corpus. As our data is just two numbers reflecting frequency of a word in two sources, we use a simple linear regression model, in particular, Ridge regression, with $\alpha = 1.0$.

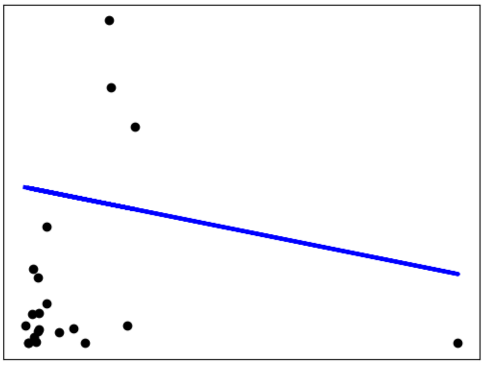


Figure 4: Regression on the basis of word frequencies in the BNC data and the VG data. The test set of 20 word types is shown.

However, the data does not fit into regression as the computed error is very high and the explained variance score is on the contrary low. An example of a built regression can be seen in picture 4.

The results for the regression models show that indeed word frequencies in the VG and the BNC differ significantly, and the initial hypothesis that words that are frequent in the real world should be frequent in the natural language does not hold. To investigate this issue in more details as well as look into the more granular parts of the vocabulary, we check the correlation scores for frequencies of objects, attributes and predicates in the VG and both corpora. To do this, we collect types, relations and attributes that are found in the VG annotation. We compute their frequencies in the model and in the BNC and English Wikipedia corpora. After this, we can simply calculate Spearman ρ correlation between two lists of frequencies. The results are presented in Table 3.

Setting	**types**	**relationss**	**attributes**
BNC	0.0199	0.3218	0.3147
Wikipedia	0.0912	0.3427	0.1676

Table 3: Correlation between frequency lists in VG with Wikipedia and BNC.

As we can see from the table, correlation between frequencies for relations and attributes is much higher than for types. This trend holds true both for Wikipedia and the BNC spaces, though attributes in Wikipedia correlate less well with the attributes in the Visual Genome. This implies that while types have different salience in both corpora, their attributes and relations they are involved in are more correlated.

Another trend that can be noticed is that the correlation between types and relations in Wikipedia and the VG is higher than in the BNC and the VG. This means that indeed the real world space (approximated by the VG) is closer to the encyclopedic language of Wikipedia than to the BNC. However, this is not true for the attributes that have higher correlation in BNC and VG. This may also mean that the language space presented by the BNC is more concerned with the attributes in general than the language space modeled by Wikipedia, where words are expressed through their co-occurrences with each other and their relations.

It is also interesting that whereas the VG models built on the basis of relations have shown the best quality as compared to the human judgments (see Section 4.3), the corresponding type frequency comparison has shown the worst result. Thus, we can conclude that although the best semantic modeling is achieved when describing concept meaning through its co-occurrence with other concepts, frequencies of concepts in the world space and the language space differ drastically and do not correlate well. At the same time, relations, that were the the best model built from the VG, show decent correlation here, both for the BNC model and Wikipedia model. Finally, attributes, that did not produce high-quality semantic representations from the VG, show lower correlation score here as well.

Another approach that we took is clustering the overlapping dimensions and computing correlation between frequencies for words from different clusters. The aim is to find out which clusters map well from one space to another in terms of frequencies and which ones do not. We have tried K-Means with manually specified number of clusters and DBSCAN (Schubert et al., 2017) to find out the number of clusters in an unsupervised way. We also performed manual clustering into four groups: human relations, areas, touchable objects, abstract objects and substances. The manual clustering seems to be biased, and the clusters are very generic. The DBSCAN clustering did not produce any meaningful results, thus, we decided on using the K-Means algorithm. The problem when using K-Means is that we need to specify the number of clusters manually, and this poses a problem when this number is unknown. We have experimented with different cluster numbers (10, 20, 50, 100) and finally used 20 clusters. Higher number of clusters makes the clusters too granular and hard to interpret. Some clusters, however, stay more heterogeneous than the other.

We have also tried to label the clusters in an unsupervised way. To do this, we extracted the immediate hypernyms for every word in the cluster. The most frequent hypernym in the cluster is then used as the label for the whole cluster. If all the immediate hypernyms have equal frequency of 1 or 2, we go one level up and extract hypernyms of the hypernyms. On this step most probably we find a probable label for the cluster. Though these labels are not ideal, in most cases they accurately reflect the semantic area of the words in the cluster. For example, the cluster *clothing* includes such words as *shirt, hair, garment, sunglasses, person, outfit.* It would be more correct to name this cluster as "visual appearance", but the given label serves well for the means of our semantic analysis.

Additionally, when looking at the clusters, we have found that they can be roughly divided into 2 types: situation-based clusters and object-based clusters. The object-based clusters seem to employ the notion of semantic similarity, as words in them are most frequently from the same semantic field and are very similar. An example of this can be the cluster *vascular plant*, which includes such words as *banana, apple, mango, coconut*, etc. Situation-based clusters seem to be closer to the notion of relatedness, as the words included in such clusters seem to be from different semantic fields and not at all similar to each other. However, they still can be imagined together in the same situation. And example of such cluster can be *motor vehicle*, which includes such words as *rear, road, truck, policeman, bumper, stoplight.* The full list of clusters with words they include and their annotated types can be found in Appendix B.

For computing correlation for types, we simply take words from every cluster, sum up their frequencies in the corpora and compute the correlation with the corresponding

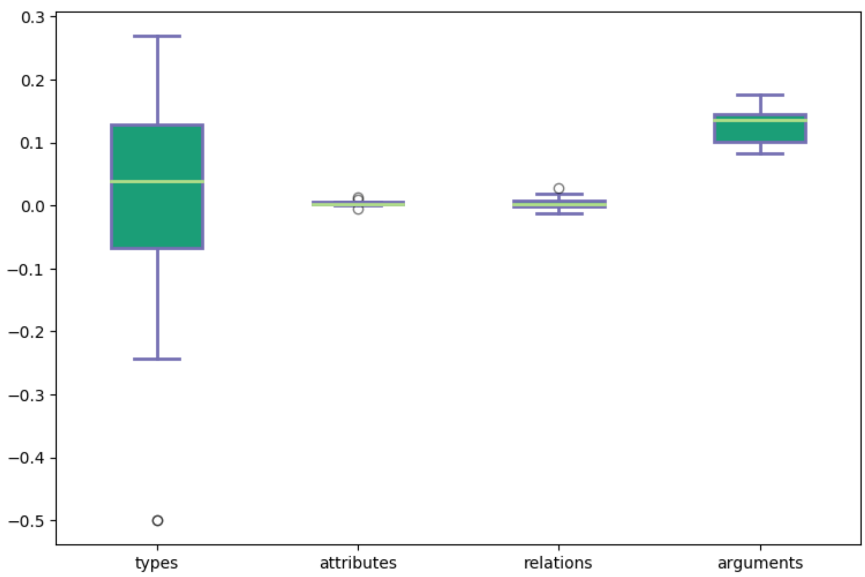


Figure 5: Dispersion of correlation scores for different settings when comparing frequencies in BNC and VG.

types in the VG annotation. Summing up the frequencies for words in the same cluster is done in order to measure the frequency of a cluster as a whole, since we are interested in how often we mention, for example, various *food concepts*, not particular exemplars (e.g. *cheese*). For computing correlation for relations and attributes, we compute co-occurrences with all possible relations and attributes for every individual word, and then concatenate these vectors for every word in a cluster. The correlation is then computed for these very long "flattened" vectors.

Another setting that we added in the analysis here is counting co-occurrences for nouns that are arguments of the same predicate. For example, from the predicate *chew(dog, ball)*, which signifies that a dog is chewing a ball, we take the co-occurrence of *dog* and *ball*. This should roughly model the *situation* setting of the VG models. The correlation numbers for different clusters and different settings can be found in the Table 4. The figures for the argument setting are given in the column *args*. The name of the clusters are given in the first column, along with the number of words included in this cluster (in brackets). The type of the cluster (situational of object-based) is given in the next column. The ordinal number of clusters are preserved for the convenience of the analysis.

The results from Table 4 are also summarized in Figures 5 and 6. On Figure 5 we can see the distribution of correlation scores based on the BNC data. The distribution of correlation scores and mean scores for frequencies in English Wikipedia is displayed on Figure 6.

The main conclusion that we induce from the scores is that the correlation between frequencies in English Wikipedia and VG or BNC and VG is quite low. Though we have correlation of 0.9 for some clusters, for the majority of data the scores stay on the level of 0.2. This tells us that the salience of concepts in the "real world model" and language models does not really correlate. However, we can try and interpret the current results concerning the differences between different clusters and different types of features.

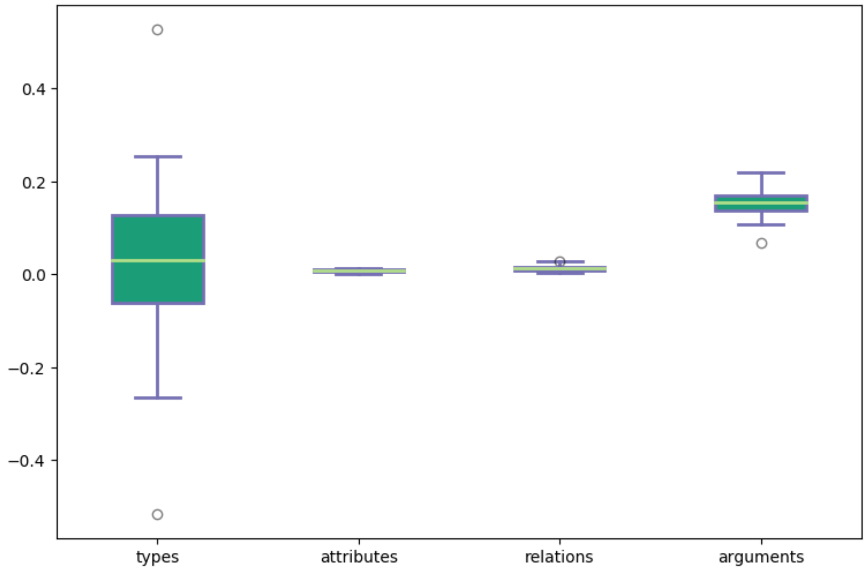


Figure 6: Dispersion of correlation scores for different settings when comparing frequencies in English Wikipedia and VG.

cluster	type	**types BNC**	**types Wiki**	**attrib. BNC**	**attrib. Wiki**	**relations BNC**	**relations Wiki**	**args BNC**	**args Wiki**
0 - natural_object (66)	situational	-0.1636	0.0197	0.0030	0.0065	0.0075	0.0106	0.1126	0.1431
1 - clothing (177)	object-based	0.0763	0.1683	0.0096	0.0112	0.0278	0.0292	0.1343	0.1465
2 - fixture (33)	object-based	-0.0344	-0.2296	0.0027	0.0036	-0.0009	0.0052	0.1642	0.1677
3 - geological_formation (39)	situational	0.1785	0.1815	0.0136	0.0133	0.0188	0.0239	0.1493	0.1618
4 - vascular_plant (18)	object-based	-0.2435	0.0481	-0.0055	0.0022	-0.0056	0.0072	0.1006	0.0670
5 - artifact (71)	situational	0.0979	0.1333	0.0029	0.0103	0.0019	0.0159	0.1760	0.1890
6 - instrumentality (270)	situational	0.1141	0.0233	0.0060	0.0102	0.0077	0.0169	0.1213	0.1246
7 - nutriment (154)	situational	0.2692	0.2553	0.0030	0.0054	0.0014	0.0126	0.1362	0.1390
8 - band (sports) (46)	situational	-0.9999	-0.2643	-0.0004	0.0066	-0.0024	0.0066	0.0836	0.1284
9 - vessel (19)	object-based	-1.0	-0.5166	0.0023	0.0005	-0.0010	0.0095	0.1443	0.1952
10 - home_appliance (37)	object-based	-0.5	-0.1811	0.0025	0.0065	-0.0015	0.0112	0.1164	0.1483
11 - geological_formation (61)	situational	0.1704	0.0666	0.0054	0.0091	0.0023	0.0121	0.0975	0.1088
12 - furniture (94)	situational	0.0861	0.0187	0.0052	0.0082	0.0040	0.0148	0.1400	0.1503
13 - facility (21)	object-based	0.0	0.0392	0.0025	0.0008	-0.012	0.0040	0.0981	0.2188
14 - device (13)	object-based	0.0	0.5270	0.0031	0.0074	0.0056	0.0141	0.1009	0.1696
15 - body_part (63)	object-based	-0.0045	0.1245	0.0038	0.0086	0.0082	0.0137	0.1446	0.1720
16 - substance (26)	situational	-0.9999	0.0420	0.0030	0.0092	-0.0001	0.0126	0.1655	0.1778
17 - artifact (160)	situational	0.2060	-0.1208	0.0063	0.0124	0.0103	0.0199	0.1438	0.1584
18 - motor_vehicle (53)	situational	-0.5	-0.0424	0.0037	0.0104	0.0108	0.0273	0.1482	0.1702
19 - plant_part (39)	situational	0.9999	-0.0416	0.0022	0.0069	-0.0026	0.0049	0.0980	0.1066

Table 4: Correlation between frequency lists in VG and BNC and Wikipedia for different clusters.

First of all, we can see that overall frequencies of types in both corpora demonstrate greater variance than other settings (attributes, relations or arguments). This corresponds to the idea that different objects can be more or less similar in language spaces and the world space, and these differences can be figured out and described separately for different semantic fields.

However, contrary to the generic correlation presented in Table 3, relations and attributes perform worse than mere types. We can hypothesize that for specific clusters the representations of nouns are different (objects are described in VG and language with different attributes and are involved in different relations). Nevertheless, their frequencies stay more or less the same both in the world and in the language. The frequencies of attributes and relations do not correlate at all. The big conclusion that can be made on this basis is that though people mention the same concepts in the language that they see in the real world, the attributes of these concepts and the relations they are involved in differ significantly. This difference is more seen when looking at different semantic clusters and cannot be noticed in the generic correlation of frequencies.

Attributes in both corpora show the lowest variance: as we can see on the figures with dispersion, most of the values oscillate around 0. Thus, co-occurrence of word types with different attributes differs entirely in the world space and the language spaces. Arguments in both corpora demonstrate highest correlation scores. This corresponds to the features of VG spaces we described earlier. Still the best way to create semantic representations can be situations – objects that occur together, or in this case arguments of the same predicate (thus, also in the same situation).

Another trend that we can find is that situation-based clusters show higher correlation scores than object-based clusters. This finding also goes in line with the idea that the best semantic representations are achieved using the situational information. This idea was also proven before in the narrative cloze evaluation studies and in the computer vision, as we pointed out in Section 2. Overall, it seems that the semantic information is better preserved in the situational information and in object co-occurrences within the same situation. This also explains why the VG models achieved higher scores on the MEN dataset than on the SimLex-999 dataset.

For better visual analysis, the results per cluster for the BNC model are shown on Figure 7. Different shades of blue correspond to correlation scores close to 0, red shows high correlation, green and yellow stands for low correlations scores. Same map, but showing types of clusters (situational or object-based), can be found on Figure 8.

Once again, for our maps we find that situational cluster exhibit high correlation scores – red color (though it is not true for all situational clusters). For object-based clusters it is much harder to achieve high correlation, and the majority of them are thus colored with blue or green.

To conclude, in this experiment we have shown that:

- Concept behaviour differs greatly in the world model and the language models, which can be seen in the low correlation scores for frequencies. People tend to talk about different concepts than those that are found in the real world.

- Words from some semantic clusters show more similarity between the world space and the language spaces. For example, correlation for frequencies for some clusters is higher than for others.

- Situational information indeed seems to be the best way to model semantic representations. Situational-based clusters have higher correlation for frequencies

Figure 7: The map of correlation values for different clusters, with unsupervised annotated labels.

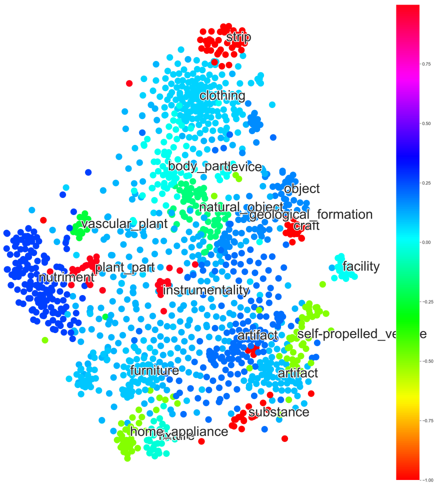
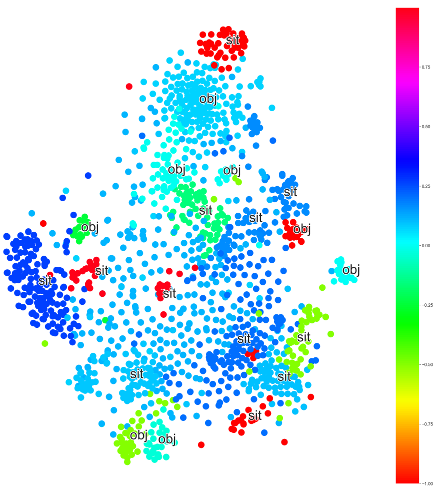


Figure 8: The map of correlation values for different clusters, with the annotation of cluster types.



with the VG data, than object-based clusters.

Overall, with this experiment we have shown that whereas the majority of concepts is represented quite differently in the world space and language spaces (which is reflected in different frequencies), some concepts do present similar behaviour as compared to the behavior of word types in the corpora. Therefore, we prove that it is impossible to directly map from a language space to the world space. However, for some categories of concepts this mapping can be more effective due to their similar behaviour across sources of data (VG vs. corpora).

5.4 Experiment 3: Comparing Neighbors

In this section we concentrate on the overlapping dimensions in both models. Previously we have studied the differences in the vocabularies of our models and the differences in word frequencies, which we interpret as salience. The purpose of the experiments in this section is to study how different are the representations of the same word in two models. Even if a word is present in both models and has approximately the same frequency, it does not mean that semantic representations are the same in all models. To perform such an analysis, we extract neighbors of a word in our spaces (VG, Wikipedia and BNC) and calculate the overlap between the neighbor lists. We test several possible sizes of neighbor lists: 5, 10, 20 and 50. We decide to use in our experiments the size 20 since it provides the most informative data yet not containing much noise. In Table 5 you can find an example of neighbor lists for a word *person* in VG, BNC and Wikipedia models.

The neighbors of the words in the models can be completely different due to huge differences in the vocabulary. First of all, corpora include much more words than the VG data, whose capacity is limited by the objects occurring in the images. Second, the VG includes in the annotation a high number of multi-word expressions, which are in their turn are not found in the corpora. To overcome these issues, as I pointed out earlier, and for the purpose of fair comparison, I construct two types of count-based models. In the first type I use all the data in the corpora without dropping any dimensions. In the second type of models I drop all the dimensions in the model that are not present in the VG data. This is done to maximize the comparability between the models. To cope with the presence of multi-word expressions in the VG, I extract 60 neighbors for each word that I analyze, remove all multi-word entities and take top 10 most similar remaining neighbors.

VG space	**BNC space**	**Wiki space**
man	man	man
woman	woman	woman
people	child	patient
girl	citizen	defendant
lady	girl	player
boy	lady	child
skier	chap	girl
guy	patient	someone
child	guy	thing
he	gentleman	person's

Table 5: Top 10 semantic neighbors of the word *"person"* in the spaces built on the VG data, BNC and Wikipedia.

As neighbor lists, we have partially intersecting sets of word associations. The degree of difference between them can be calculated using the Jaccard similarity coefficient (Jaccard, 1901; Niwattanakul et al., 2013; Choi et al., 2010), which is calculated as follows: $J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$ (the size of the intersection divided by the size of the union of two sets). It takes values in the interval [0,1] and serves here as a measure of how much the meaning of a given word in one model is different from the meaning in another model. If the Jaccard coefficient equals to 0, that means that the two sets of neighbors do not intersect and thus the meaning is supposed to be totally different. If, on the contrary, the coefficient takes the value 1, the two sets are identical: both models provided precisely the same set of 10 nearest semantic neighbors.

First, we will talk about comparing neighbor lists in the full corpora models, where the dimensions absent in the VG data were preserved. We calculated the overlap for 100 most frequent concepts in the Visual Genome data to exclude concepts for which good representations were not learned due to the sparsity of data. The Jaccard coefficient ranges from 0.333 for the most overlapping neighbor lists to 0.0 for most words in the list of comparison. In fact, there are only 20 words for which the coefficient is higher than 0.1, and for 50 words the coefficient equals to 0. It means that indeed as we pointed out earlier, the representations of concepts in the language and in the truth-theoretic model are entirely different, since the concepts have different distributional properties.

The concepts that overlap most in the BNC corpus and in the VG model are the ones that pertain to humans, their appearance and body parts. You can see some examples of such concepts in Table 6. Among such words we can see *face, human, woman, child,* *head, person, female*. Their semantic neighbors in both corpora models are roughly the same: another body parts for body part concepts and another human roles for concepts like *human, woman, child*. This means that the contexts in which we see such words are uniform in our models. It is interesting to notice that high overlap is achieved also in the cases of homonymy, like for the word *head*. Another category of words with high overlap is location names: *street, office, building, town, street*. It is interesting to notice that for the VG data these are very generic concepts and most often describe the whole set of objects situated on the image. Therefore, it seems that in the natural language this generic name places are also described using the components of the situations they are met in.

Concept	**Jaccard coefficient**	**VG neighbors**	**BNC neighbors**
face	0.333	eye ear head nose mouth legs eyes chin ears	brow forehead nose eye chest belly mouth chin head
person	0.290	man woman people girl lady boy skier guy child	man woman patient defendant player child thing someone person's
child	0.290	girl boy baby kid woman person skier adult lady	woman baby man person girl patient mother daughter victim
woman	0.250	lady girl man person boy child female guy she	man girl boy child person baby soldier doctor priest
man	0.212	person woman boy guy girl lady he male child	woman chap boy girl person gentleman guy lad soldier
head	0.212	eye ear captain nose mouth face legs tail chief	chairman secretary chair seat governor commander face captain chief

Table 6: Top 9 semantic neighbors of the words with most overlap in the neighbor lists in the spaces built on the VG data and BNC.

Concepts that overlap least in the BNC and the VG include more abstract ones as well as more detailed human roles and more detailed locations. We can find with the lowest Jaccard coefficient such words as *bank, bill, business, court, employee, customer, speaker*. The representations for these objects are not good in the VG despite that they were frequent enough to be included in the analysis. There are also cases of severe homonymy, where one word sense prevails in the VG and another in the BNC, e.g., the words *bank, court*. The neighbors of human roles in the VG are other human roles, whereas in BNC they are terms pertaining to the same lexical field (e.g. economic terms). You can see several examples of the least overlapping words and their neighbors in the Table 7.

When we look at the clusters for the words with the highest and the lowest Jaccard coefficients, we can notice that words from object-based clusters tend to score higher than the words from the situation-based clusters. This can be explained by the fact

Concept	Jaccard coefficient	VG neighbors	BNC neighbors
bank	0.0	shoreline seawall coast mountian shore mounds foilage tents cliff	banks deposit canal company river cheque fund lake taxpayer
court	0.0	pitch field airstrip running pavement road turf asphalt net	trial courts defendant parliament judgment plea council request judicial
customer	0.0	register melons patron dad wineglass gazebo barber biker stools	client buyer purchaser seller user patient supplier person debtor
employee	0.0	thing materials grassland workers fair timber tarmac dolly halfpipe	agent assistant attorney expert officer owner editorial opponent indication
object	0.0	objects radio linen baseboard things shells stuff items magazines	utterance organism entity instrument attribute image objection objective illusion
speaker	0.0	speakers mouse printer cords monitor video console keyboard sports	researcher writer poet operator reader critic composer priest policeman

Table 7: Top 9 semantic neighbors of the words with least overlap in the neighbor lists in the spaces built on the VG data and BNC.

that situations that require more complex descriptions in the language therefore provide more possibilities for differences between the models. In their turn, the object-based concepts are easier to describe with annotations in the images, and the annotations are more likely to coincide with natural language descriptions for the objects. However, it contradicts the findings from the previous sections, where we figured out that words from the situation-based clusters have higher correlation scores for the frequencies, which means that they are equally salient in the natural language and the real world.

When we try to reduce the BNC space and drop all the dimensions that are not present in the VG, we do not observe much change in the results. In fact, the words with the highest Jaccard coefficient and with the lowest Jaccard coefficient stay the same, while only for a couple of words their position in the lists drastically changes. The majority of semantic neighbors for all the concepts stays the same in the reduced space. Thus, we can conclude that the important semantic information stays in the reduced space as well.

Overall, from the analysis of neighbor lists for concepts in our models we can make the following conclusions:

- the overlap between neighbor lists for the majority of concepts in the BNC and the VG is very low (the highest Jaccard coefficient is only 0.333, one third of the semantic neighbors is the same in both models). This tells us that indeed the distributional properties of words in both models are very different, even for the words with high frequency correlation.

- Words with the highest overlap include terms describing people (*man, woman, child, person, female*), body parts (*face, head, hand, leg*) and locations and buildings (*sea, park, street, town*). These are the words that own approximately the same distributional properties both in the VG and the BNC.

- Words from the object-based clusters tend to have higher overlap of semantic neighbors than the words from the situaion-based clusters. Thus, people use approximately the same lexical units to describe objects in the real world and the natural language, but the descriptions of situations may be different.

- We are also interested in the differences in neighbor lists between nouns with high overlap and low overlap. We have already mentioned which entities have high overlap in neighbor lists between spaces – they mostly depict concrete objects from everyday life and are highly frequent both in the corpus and in the VG data. Among the entities that have low overlap in the neighbor lists are areas such as *forest* or *village*.

Overall, this experiment has shown us that there is indeed a major distinction between situational information about the concepts (i.e. relatedness notion) and information about attributes of concepts (i.e. similarity notion). As it was noted in the evaluation of the VG space, the relatedness notion seems to be better captured when creating a semantic space. Therefore, relations and situations proved to be the best model architectures for the VG models. In this experiment we see that situational information is also much more stable when comparing across the models, whereas the relatedness notion, expressed in object-based clusters, proves to be much less stable in different semantic models.

Also, in this experiment we have identified which lexical clusters have more potential to be mapped across world-based and corpus-based semantic models. Whereas this direction of work needs deeper investigation and particularization, we already can hypothesize which lexical categories are more likely to be corresponding when mapping our models.

5.5 Discussion

As we have described our initial hypothesis in Introduction, in this research we were trying to figure out whether the semantic space created on the basis of the real world data corresponds to the semantic space created on the basis of the corpus data. This correspondence can be measured, first of all, through evaluating both spaces with regard to the behavioural data. Our research has shown that both types of spaces correlate well with behavioural data and, thus, semantic representations from both spaces agree with human perception. Nevertheless, when it comes to comparing the spaces directly, we find out that the representations in the spaces differ fundamentally, as they are shaped differently in the human conceptual space. The situational information and relation between concepts seem to be more stable across the spaces, and attributes of concepts differ drastically between spaces. Overall, our results are described on Figure 9, which shows how our experiments have resolved the questions posed on Figure 1.

We have also named possible reasons why our language space differs from the real world space can be:

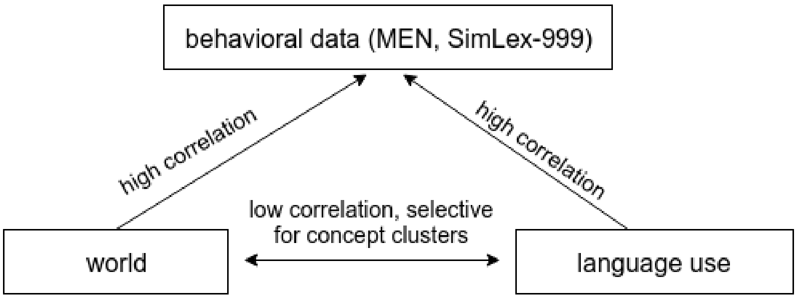

behavioral data (MEN, SimLex-999)
high correlation
high correlation
low correlation, selective for concept clusters
world
language use

Figure 9: A scheme of the research conclusions.

- we don't talk about the world as it is because we describe all things in it through the prism of our perception (prototype theory or attention mechanisms (Rosch et al., 1976));

- we don't talk about the world as it is because we're Gricean and thus do not want to talk about obvious and unimportant things (Grice et al., 1975);

- we don't talk about the world as it is because of human creativity, as we create other possible worlds with our language. This hypothesis remains for further research, since we didn't dive into creative language in our project.

It seems that each one of these reasons contributes to the semantic shifts between spaces. Perception is influencing every property of concepts in our models. We have seen how differently the concepts are named in the world-based space because giving more concrete names to objects does not seem to be vital in a particular situation. The second reason was investigated in more details in the work on the color subspace (Rawee, 2018), but in our research we have also seen its signs, for example, when we find out that people are less likely to name a person's profession, possibly out of fear to provide false or redundant information. The third reason is harder to investigate since the Visual Genome data is much more scarce than the data from the corpora. However, we can hypothesize that the human creativity is the reason why similarity notion is much less stable across models than the relatedness notion. Thus, the relations and situations in which objects are involved do not change that much when shifting from the real world to the language. However, the object characteristics expressed with attributes differ significantly. In fact, in the language the object possess much more various and detailed attributes.

The overall conclusions of the research can be described in the following way: we do find significant differences between concept representations in the world-based space and in the corpus-based spaces. In general, the majority of concepts seems to show different properties across spaces. The concepts do not correspond in terms of their frequency, or semantic neighbors. However, some particular categories of concepts differ less significantly between spaces than other categories. Describing such categories in every detail can be possible extension of the present work.

6 Conclusion

In the present research we tried to explore how different are semantic representations of the world create from truth-theoretic data (annotated images) and extracted from language data. The implications of our work lie in several directions.

The first direction is an attempt to unify tools of formal and distributional semantic. Both these branches of semantics have their own advantages and disadvantages, but their unification provides a really powerful tool for studying the interaction between similarity and relatedness, as well as finding out which properties human tap into when making association judgments. This direction of work was fully explored in the first of our experiments 4. We have shown that we can study the distributional behaviour of concepts from a large truth-theoretic model. Thus, standard distributional semantics is not unique in accounting for conceptual distance.

Further, the vector spaces we created have the advantages of formal models, by linking to a clear notion of entity and associated properties. We have also demonstrated that by choosing the right properties, the truth-theoretic vector space achieves superiour performance compared to a usage-based DSM on considerably less data. This is probably the most important finding of our first experiment, because we found out that the quality of created models differs significantly when we take different features into account. This finding suggests that not all properties that we can use to construct meaning representations contribute equally to the final quality of a model. We tried to further investigate these differences in our next experiments. While this point does not have practical application, we believe this result may have implications for understanding how humans themselves build concepts from the limited set of situations they are exposed to.

We have also found out that though in general the similarity between semantic representations of concepts between the language model and the world model is very low, indeed for some concepts the difference is less striking than for the others. We have found out that the categories of concepts with higher similarity can be used to describe situations rather than concrete objects. One very salient example is the cluster of "marina" words: *boat, dock, sail, pier, float, anchor, bay*. The frequencies of these concepts were correlated both in the VG and BNC, and their neighbor lists were also overlapping. From such cases we conclude that the relatedness notion between concepts modeled both in visual and language data is similar. At the same time, categories of concepts that describe concrete things (food, appearance) differ more significantly between models. This finding shows that the widely-accepted approach of building a general semantic model for studying language phenomena without any differentiation is not appropriate.

Another direction of our work was contrasting models build on English Wikipedia and BNC when comparing them to the VG model. We have shown that semantic properties of concepts as they are built on the basis of English Wikipedia are closer to the properties of the corresponding concepts in the VG. This is shown in the lower dispersion of correlation score for frequency lists in Wikipedia and VG. Thus, we can conclude that encyclopedic knowledge contained in Wikipedia better correlates with object properties in the real world.

In future, we plan to correct some limitations of our current evaluation. In particular, our VG models are not fully comparable with standard count-based spaces, as different pairs from the datasets Simlex-999 and MEN are evaluated (the overlap between pairs evaluated for text spaces and VG spaces is on average 50%).

We also plan to apply more comparison methods to reveal more fine-grained differences between spaces. The next direction of our research is to create a mapping function between spaces to understand what transformation a concept representation undergoes when being mapped from VG to BNC.

Another direction for improving our results is to experiment with other image-annotated datasets or knowledge graphs to further understand which formal relations might be at the basis of human similarity judgments. Indeed, one problem with our current setup is that by using vision data, we limit our evaluation to concrete concepts.

References

Andrews, M., G. Vigliocco, and D. Vinson (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological review 116*(3), 463.

Asher, N., T. Van de Cruys, A. Bride, and M. Abrusán (2016). Integrating type theory and distributional semantics: a case study on adjective–noun compositions. *Computational Linguistics 42*(4), 703–725.

Baroni, M., R. Bernardi, N.-Q. Do, and C.-c. 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*, pp. 23–32. Association for Computational Linguistics.

Baroni, M., G. Dinu, and G. Kruszewski (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pp. 238–247.

Baroni, M. and A. Lenci (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics 36*(4), 673–721.

Beltagy, I., C. Chau, G. Boleda, D. Garrette, K. Erk, and R. Mooney (2013). Montague meets markov: Deep semantics with probabilistic logical form. In *Second Joint Conference on Lexical and Computational Semantics (*SEM2013)*, Atlanta, Georgia, USA, pp. 11–21.

Bengio, Y., R. Ducharme, and P. Vincent (2003). A neural probabilistic language model. *Journal of Machine Learning Research 3*, 1137–1155.

Bernardi, R., G. Dinu, M. Marelli, and M. 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)*, Volume 2, pp. 53–57.

Boleda, G., M. Baroni, T. N. Pham, and L. McNally (2013). Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS2013)*, Potsdam, Germany, pp. 35–46.

Boleda, G. and A. Herbelot (2016). Formal distributional semantics: Introduction to the special issue. *Computational Linguistics 42*(4), 619–635.

Bruni, E., G. Boleda, M. Baroni, and N.-K. Tran (2012). Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 136–145. Association for Computational Linguistics.

Choi, S.-S., S.-H. Cha, and C. C. Tappert (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics 8*(1), 43–48.

Clark, S. (2012). Vector space models of lexical meaning. In S. Lappin and C. Fox (Eds.), *Handbook of Contemporary Semantics – second edition*. Wiley-Blackwell.

Coecke, B., M. Sadrzadeh, and S. Clark (2011). Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis: A Festschrift for Joachim Lambek 36*(1–4), 345–384.

Emerson, G. and A. Copestake (2016). Functional distributional semantics. *arXiv preprint arXiv:1606.08003*.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: a survey. *Language and Linguistics Compass 6*, 635–653.

Erk, K. (2016a). What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics 9*, 17–1.

Erk, K. (2016b). What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics 9*, 17–1.

Fagarasan, L., E. M. Vecchi, and S. Clark (2015). From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *Proceedings of the 11th International Conference on Computational Semantics*, pp. 52–57.

Faruqui, M. and C. Dyer (2015). Non-distributional word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL2015)*, Volume 2, pp. 464–469.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. studies in linguistic analysis. *Oxford: Philological Society. [Reprinted in Selected Papers of J.R. Firth 1952-1959, ed. Frank R. Palmer, 1968. London: Longman]*.

Garrette, D., K. Erk, and R. Mooney (2011). Integrating logical representations with probabilistic information using Markov logic. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS2011)*, pp. 105–114.

Grefenstette, E., M. Sadrzadeh, S. Clark, B. Coecke, and S. Pulman (2014). Concrete sentence spaces for compositional distributional models of meaning. In *Computing meaning*, pp. 71–86. Springer.

Grice, H. P., P. Cole, J. Morgan, et al. (1975). Logic and conversation. *1975*, 41–58.

Gupta, A., G. Boleda, M. Baroni, and S. Padó (2015). Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 12–21.

Herbelot, A. (2013). What is in a text, what isn't, and what this has to do with lexical semantics. In *Proceedings of the Tenth International Conference on Computational Semantics (IWCS 2013)*, Potsdam, Germany.

Herbelot, A. and E. M. 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*, pp. 22–32.

Hermann, K. M. and P. Blunsom (2013). The role of syntax in vector space models of compositional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Volume 1, pp. 894–904.

Hill, F., R. Reichart, and A. Korhonen (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics 41*(4), 665–695.

Hürlimann, M. and J. Bos (2016). Combining lexical and spatial knowledge to predict spatial relations between objects in images. In *Proceedings of the 5th Workshop on Vision and Language*, Berlin, Germany, pp. 10–18.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat 37*, 547–579.

Krishna, R., Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision 123*(1), 32–73.

Kuzmenko, E. and A. Herbelot (2019). Distributional semantics in the real world: building word vector representations from a truth-theoretic model. In *Proceedings of the 13th International Conference on Computational Semantics-Short Papers*, pp. 16–23.

Larsson, S. (2013). Formal semantics for perceptual classification. *Journal of logic and computation 25*(2), 335–369.

Lazaridou, A., E. Bruni, and M. Baroni (2014). Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1403–1414.

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

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

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119.

Mikolov, T., W.-t. Yih, and G. Zweig (2013). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pp. 746–751.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography 3*(4), 235–244.

Niwattanakul, S., J. Singthongchai, E. Naenudorn, and S. Wanapu (2013). Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, Volume 1, pp. 380–384.

Padó, S. and M. Lapata (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics 33*(2), 161–199.

Rawee, J. (2018). The color subspace in distributional semantics: Between utterance conservation and world transformation. *Master's thesis, University of Trento*.

Recski, G., E. Iklódi, K. Pajkossy, and A. Kornai (2016). Measuring semantic similarity of words using concept networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pp. 193–200.

Rosch, E. (1999). Principles of categorization. *Concepts: core readings 189*.

Rosch, E., C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem (1976). Basic objects in natural categories. *Cognitive psychology 8*(3), 382–439.

Rubinstein, D., E. Levi, R. Schwartz, and A. Rappoport (2015). How well do distributional models capture different types of semantic knowledge? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 726–730.

Sahlgren, M. and A. Lenci (2016). The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 975–980.

Schubert, E., J. Sander, M. Ester, H. P. Kriegel, and X. Xu (2017). Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS) 42*(3), 19.

Turney, P., P. Pantel, et al. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research 37*(1), 141–188.

Turney, P. D., Y. Neuman, D. Assaf, and Y. Cohen (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 680–690. Association for Computational Linguistics.

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

Vecchi, E. M., M. Baroni, and R. Zamparelli (2011). (linear) maps of the impossible: capturing semantic anomalies in distributional space. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pp. 1–9. Association for Computational Linguistics.

Wang, S., S. Roller, and K. Erk (2017a). Distributional modeling on a diet: One-shot word learning from text only. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, Volume 1, pp. 204–213.

Wang, S., S. Roller, and K. Erk (2017b). Distributional modeling on a diet: One-shot word learning from text only. *arXiv preprint arXiv:1704.04550*.

Westera, M. and G. Boleda (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.

Young, P., A. Lai, M. Hodosh, and J. Hockenmaier (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics 2*, 67–78.

A List of Non-Overlapping Dimensions in VG and BNC

Below you can find a list of non-overlapping concepts in the VG dataset and the BNC corpus that were analyzed in Section 5.2. All the words are given in the lower case.

transcription
geography
clerk
father
poet
historian
deputy
theatre
member
emperor
instrument
community
nurse
department
copy
product
publishing

kingdom
doctor
assistant
message
list
sociologist
metropolitan
afternoon
facility
residence
centre
countryside
chamber
region
aunt
peasant
stomach

item
earl
republic
district
equipment
valley
author
node
bishop
politician
song
wife
mile
press
visit
page
victim

governor
moon
conservation
dad
discus
card
distal
morning
worker
coast
flesh
daughter
rat
resident
client
episode
clinic
lip
temple
institute
miner
uncle
club
prisoner
brain
muscle
cash
closing
lawyer
county
brother

gene
visitor
paragraph
diet
chief
council
sister
device
cancer
university
project
tear
weapon
chapter
junction
site
interview
citizen
mine
channel
eat
drive
north
map
teacher
conference
bowel
receiver
protein
throat
saw

league
husband
east
partner
zone
tongue
stock
root
hall
sound
contact
south
voice
colleague
network
agent
resource
west
miller
queen
village
oxygen
jew
town
secretary
draft
duke
sodium
library
calcium

B List of Clusters

In this Appendix you can find a list of clustered concepts together with annotated types and unsupervised labels for each cluster.

Cluster	Label	Num.	Type	Concepts
0	natural_object	66	situational	leg grass tail field trunk branch soil neck giraffe zebra animal horn hoof spot mane body weed ivory log vegetation branchlet hay abdomen boulder enclosure stomach moss bark snout goat menagerie mud wool pasture lamb stump bulge barn dirt cattle herd calf butt plain stream tuft feeder farm mother udder deer hayfield rear underbelly antler wild antelope wildflower trough foreleg droppings gazelle exhibit habitat buttocks brook
1	clothing	177	object-based	man shirt person hand hair woman people trouser arm jacket hat foot helmet logo jean girl cap spectacles glove coat sunglasses skateboard boot necktie sleeve lady strap suit backpack wrist knee dress watch guy sweater air camera frisbee belt vest jersey set shoulder bag pocket skirt ramp watchband apparel scarf necklace beard group hood sandal elbow tent outfit accountant crowd clothing couple lace monocle ponytail pitcher tarpaulin mustache tattoo sweatshirt apron restaurant skateboarder flop blouse pair earphone cooler clasp shoelace lap male helium waist microphone short female battalion bandanna saunterer heel adult graphic beanie sole wristwatch lens ankle family guard costume cigarette badge down hip lanyard guitar brace skater trick gauze singlet robe tungsten photographer blazer braid chalk goatee legging mesh bib briefcase brim slipper stocking overall lapel gun phase cane friend brace blackboard padding spectacle toddler cape skate slump forearm tripod standing drum dais seating sideburn tartan handkerchief decoration cardigan sphere customer tights gown fist turtleneck capri dandy limelight parade parka cafe dad gentleman swing ma insignia sitting event eagle blond raincoat haircut wheelchair brassiere skateboarding

2	fixture	33	object-based	tile towel sink eyelid toilet faucet pipe bathroom holder bathtub tank drain shower soap tissue dispenser urinal grout basin stall valve toiletry pat toothpaste lever speculator dryer plumbing toilet chrome lotion pipe shampoo
3	geological_formation	39	situational	cloud water wave sand surfboard beach ocean ripple shore foam surfer footprint land horizon splash sea island shoreline cliff paddle spray swimsuit bikini aftermath seaweed crest parasail sunset surf silhouette raft parachute boardwalk surfing seashore dune tide mist wind
4	vascular_plant	18	object-based	banana fruit apple bunch crate chaff pineapple pear market cardboard peel scale produce mango bruise melon plantain coconut
5	artifact	71	situational	building sidewalk street roof tower curb bridge van balcony awning shop chimney streetlight city signal fireplug steeple station crossing entrance billboard pedestrian crane traffic railway intersection dome garage church walk skyscraper cab lane manhole degree roadway shopfront lamppost manner median business town overpass shelter gutter mailbox sedan hotel minivan night parking overhang moon streetcar apartment rooftop dumpster weathervane palace metro pavilion facade scaffolding narrative transformer court alley finish bend vane crossing
12	furniture	94	situational	wall chair floor picture pillow shelf lamp frame curtain ceiling rug bed base sofa room cord desk laptop blanket screen keyboard television shade computer blind monitor mouse sheet speaker painting fixture headboard control stool quilt lampshade magazine fireplace art bedroom bookshelf bedspread booklet chandelier mattress artwork furniture wallpaper cadmium bookcase notebook candlestick sill radiator ottoman mantle hallway molding office printer periphery paneling mouse calendar fire heater mousepad radio charger armchair desktop envelope tablet pajama bedclothes piano ashtray portrayal cornice footboard electronics case mantel stereo garment molding thermostat linen notepad affray document stapler urn drive

6	instrumentality	270	situational	leaf rock band bag bottle umbrella plant brick button part bird box board dog edge book child contemplation kite paper seat elephant cow sheep design bear photograph wood hole barroom telephone basket pot fabric key object base circle cone stick rope candle bottom label tip string chain item feather bag surface thunderbolt can beak point mark screen panel crack square decoration star baggage pen form poster bucket fan block nail print ribbon pane case plaything slat teddy section blade end brush pad pile center mat stain bin spoke image duck five rod pond scissors toothbrush magnet tape balloon hook plastic bulb heart material machine package pebble hosiery marker barrel opening lock toe debris ice bell stack drawing card triangle newspaper note bubble diamond fish inside device tube berry cage pigeon bull light tool decal outdoors glare gull bead figurine speck hanger space shape armrest slab latch ridge bristle fire leave pin pencil vessel jug fold videodisk homo date compartment sack wrap flap abrasion streak wallet cringle stud groove butterfly slot skull thing accent binder merchandise bill leather dash vein thread goose drop cast ingot gap dust circular tassel spike snowflake lining droplet shell tin flamingo palette detail movie lion ipod coin desert monkey sewing marble cactus small spout liner flower outline broom footfall parasol supply cut drop frond port stud link swan kayak bone nozzle narration rubber blood binding bobbin gauge battery remainder barge storage cube pearl rib check spot rainbow hoop smuggler back slit rung smudge alcohol soldier cartoon there turkey dinosaur dent headrest penguin signature pew owl edging check nothing almond pottery rafter easel round radiance skeleton pouch
18	motor_vehicle	53	situational	car wheel road tire bicycle mirror headlight bus truck motorcycle windshield vehicle pavement wiper bumper grillroom handlebar meter driver batch asphalt reflector fender dawdler cab minibike stoplight kickstand driveway policeman hubcap rear taillight highway exhaust moped cyclist jeep wagon silencer wayside brake pickup tractor license bull motorcyclist blinker police cycle splashboard camper buggy

19	plant_part	39	situational	flower vase root doughnut cake petal frosting ring rotter scattering frost cupcake pastry cookie nut candy sugar bud bouquet muffin bagel root layer glaze powder blooming good bride sunflower whirl tulip daisy elf groom grade case agreement doily dainty
7	nutriment	154	situational	table plate glass food bowl cup container pizza cheese fork vegetable knife carrot napkin tomato bread broccoli tray dish sauce spoon meat onion pepper piece crust sandwich slice wine pan topping jar boodle bun utensil potato pepperoni tablecloth mushroom straw liquid egg chump drink chicken bean strawberry beer disk salad lemon rice seed sausage bit coffee catsup cucumber cream menu bacon juice pickle spinach dessert hotdog silverware meal ham noodle spice negligee shaker beverage corn blueberry pea soup spatula carton herb condiment butter one-half cabbage milk foil coaster oil toothpick pie basil ball saltshaker tine flavorer runt toast salt cauliflower grain pancake prong dough chopstick pasta tongs chef beef squash garnish parsley tea tabletop grain radish dinner syrup floret bite burger breakfast cracker flour garlic lunch gravy grease zucchini dressing jam asparagus cayenne gelatin bite mayonnaise biscuit gusto wedge leek steak cork ingredient reception beet hamburger chili filling macaroni coriander blackberry raisin eccentric mozzarella
8	band (sports)	46	situational	shoe sock ball player racket uniform bat court spectator catcher internet ballplayer batter baseball wristband umpire cleat mask game headband mound visor tennis wristlet bleacher dugout stadium audience sweatband turf clay scoreboard team coach goal outfield match armband baseline referee avocation athlete pitch nike armlet soccer
9	vessel	19	object-based	boat dock sail mast buoy ship sailboat pier cabin oar motor seaport canoe beacon duct float bay anchor marina
10	home_appliance	37	object-based	handle cabinet counter knob drawer rack kitchen stove oven switch microwave hinge refrigerator burner countertop cupboard plug dial dishwasher appliance kettle socket blender baseboard toaster doorknob deep-freeze scope rag teapot steel sponge pump sociable pull hardware timer

11	geological_formation	61	situational	tree sky land shadow fence shrub snow background mountain hill ski area distance goggles skier park sun way snowboard slope river lake hillside lawn trail limb yard day limb snowboarder foreground bank landscape pool garden greenery shrubbery lift extremum shed valley range terrain fog snowsuit pine skiing lodger mulch copyright hut slide mountainside mitten weather google gazebo winter shovel riverbank leap
13	facility	21	object-based	airplane wing engine propeller cockpit track smoke jet airport gear tarmacadam aircraft pilot stabilizer terminal fuselage turbine contrail airdock formation missile
14	device	13	object-based	horse saddle harness rider bridle rein stirrup jockey reign pony halter winker cowboy
15	body_part	63	object-based	head ear eye nose face mouth finger cat paw collar tag fur ring tooth hair's-breadth earring eyebrow thumb lip baby thorax skin nostril smile wrinkle tongue claw chin bow leash fingernail brow knot doll buckle cheek thigh cuff knock eyelash seam student torso breakwater kitten breast toenail lipstick knuckle coil freckle jaw expression bamboo talon eyeball comb bandage puppy strand navel index lash
16	substance	26	situational	sign letter word inscription arrow text bracket stop second phosphorus thyroxine oxygen page information radius bacillus nitrogen liter direction signpost direction gram title iodine signboard hydrogen

17	artifact	160	situational	window pole light line door numeral post path bench light clock train railing flag house writing wire back side corner paint banner platform measure ad pillar rubbish gravel trim name column step gate concrete graffito statue doorway palm row cable vent rim walk cart structure symbol marker scene sunlight display barrier ladder arch cement shutter strip ledge canopy planter passenger color cross emblem garbage antenna outside stairway support pool deck hedge grate two arch figure home partition earth porch equipment vine position patio sculpture plaque gray bannister three point shingle palm acid ten tunnel covering ashcan rain set siding steam booth open one day roadblock rectangle map four worker lantern wreath fountain crown six mannequin curve ivy large twelve shield locomotive litter iron windowsill carving skylight conductor pylon index cargo nine canvas eight cobble seven booth mural memorial tourist cylinder smokestack connection year wording hovel carrier angel wicket foundation flowerbed dragon galley address employee flat spring museum exit sword plaza