Concepts and properties in word Spaces*

Marco Baroni & Alessandro Lenci

Properties play a central role in most theories of conceptual knowledge. Since computational models derived from word co-occurrence statistics have been claimed to provide a natural basis for semantic representations, the question arises of whether such models are capable of producing reasonable property-based descriptions of concepts, and whether these descriptions are similar to those elicited from humans. This article presents a qualitative analysis of the properties generated by humans in two different settings, as well as those produced, for the same concepts, by two computational models. In order to find high-level generalizations, the analysis is conducted in terms of *property types*, i.e., categorizing properties into classes such as *functional* and *taxonomic* properties. We discover that differences and similarities among models cut across the human/computational distinction, suggesting on the one hand caution in making broad generalizations, e.g., about "grounded" and "amodal" approaches, and, on the other, that different models might reveal different facets of meaning, and thus they should rather be integrated than seen as rival ways to get at the same information.

1. Introduction

The notion of *property* plays a central role in cognitive science and linguistics. Apart from the proponents of "conceptual atomism" (Fodur 1998), a larger consensus exists around the idea that concepts and meanings are complex assemblies of properties or features. Various behavioral tasks concerning semantic memory (e.g., categorization, similarity, inference, etc.) are modeled as processing at the level of the properties that compose concepts. Properties are themselves bits of conceptual structures, and their cognitive status and organization is at the center of a wide debate (Salomon & Barsalou 2001; Vigliocco & Vinson 2007). Independently of the specific form in which we can represent properties (feature lists, semantic networks, frames, etc.), a key issue is exactly how to capture the very notion of *being a property of a concept*. A possible answer to this question

Rivista di Linguistica 20.1 (2008), pp. 53-86 *(ricevuto nell'ottobre 2008)*

is that properties are salient aspects or attributes associated with or shared by a category of entities, which enter into the constitution of the concept for that category. According to this view, the fact that a particular feature (color, shape, behavior, action, etc.) is typically observed co-occurring with a certain category of entities is strongly related to its becoming one of the properties that form the conceptual representation of the category.

Concepts and properties surface in language as words and phrases, and they provide a semantic interpretation for these linguistic elements. Through language, fragments of our conceptual structures are communicated to other speakers, in turn influencing their knowledge of the world. A long-standing tradition has pointed out the key role played by the way words distribute in texts and co-occur with other linguistic expressions in shaping their semantic content. More recently, the hypothesis that corpus-derived word co-occurrence statistics can provide a natural basis for semantic representations has also been gaining growing attention in cognitive science (Landauer & Dumais 1997; Vigliocco & Vinson 2007). Some variation of the so-called 'distributional hypothesis' – i.e., words with similar distributional properties have similar semantic properties – lies at the heart of a number of computational approaches commonly known as "word space models" (Sahlgren 2006). These share the assumption that it is possible to represent the semantic content of words in vector spaces built through the statistical analysis of the contexts in which words co-occur. Distributional models of meaning are directly related to the classical discovery procedure of the structuralist tradition (Harris 1968) and to the collocational analysis typical of corpus linguistics (Firth 1957). Both have gained new momentum thanks to the availability of large-scale textual corpora, access to more sophisticated mathematical techniques to model word statistical co-occurrence, and – last but not least – the development in the last decades of an infrastructure for the computerized analysis of linguistic data that has turned the distributional approach into an effective tool for building lexico-semantic representations from texts.

A major question concerns the relationship between concepts and semantic representations as clusters of properties on the one hand, and as corpus-based co-occurrence distributions on the other. More specifically, we intend to clarify to what extent the linguistic expressions that more significantly co-occur with a word are correlated with the properties that human subjects typically ascribe to the concept expressed by that word. The long-term aim of this research is to achieve a better understanding of the relation between the notion of

property of a concept as a cognitive construct, and the semantic properties of a word as determined by its syntagmatic and paradigmatic distribution.

Investigating these issues is an essential step towards an effective evaluation of the potential for corpus-based distributional representations to be taken as models for the human property space, as well as towards a real understanding of the type of semantic information word space models are able to provide. Although distributional models have been proposed as plausible simulations of human semantic space organization, careful and extensive explorations of such claim are still lacking, with few notable exceptions such as Vigliocco et al. 2004.

With this goal in mind, we will carry out an in-depth comparison between corpus-based property spaces generated by distributional models and subjects' elicited property spaces. Two highly different types of human property spaces (Section 2) will be compared to two different approaches to semantic modeling based on distributional data extracted from corpora (Section 3). The four target spaces will be analyzed in terms of the type of properties associated with different semantic classes of concepts expressed by concrete nouns (Section 4). This multi-way analysis allows us to look at similarities and differences both within human and computational models, and between these two categories.

As far as we know, we are the first to propose a qualitative comparison of human and computational spaces in terms of property types (the computational literature, in particular, has focused on objective measures of performance, but very little work has been done on the analysis of why the models behave the way they do). Moreover, we introduce the new StruDEL word space model (Section 3.2), and we might be the first to look at the ESP Game data (Section 2.2) from the point of cognitive science.

2. Human property spaces

Many researches in cognitive psychology have recognized the added value provided by 'property generation' tasks as a source of evidence to achieve a better understanding of the human property space, i.e., the features that compose the structure of concepts. In these tasks, subjects are typically presented with a concept name and are asked to generate the properties they consider important in order to describe the concept. The elicited data are then collected

into 'semantic feature norms', i.e., list of concepts with the properties most frequently produced by subjects in response to a set of target concepts.

Feature norms are used to test the predictions of theoretical and computational models of human semantic memory. For instance, Wu & Barsalou (in press) used a property generation task to compare theories of concepts based on perceptual symbol systems versus those that assume amodal properties. Moreover, feature norms have been used to construct stimuli for further experimental research on semantic priming (Vigliocco et al. 2004), property verification (Cree et al. 2006), semantic category specific deficits (Vinson et al. 2003), etc.

Many psychologists warn against a literal interpretation of semantic norms as if they were "snapshots" of the property structure of concepts (McRae et al. 2005). However, as long as subjects use their semantic representations when asked to generate properties for a concept, these data can be used as important probes to investigate the organization of human semantic knowledge. For instance, they can provide information about the 'type' of properties generated by the subjects, their degree of distinctiveness, as well as the correlation between property types and different semantic categories. Some of these issues will also be touched in the analysis that we will present in Section 4.

Subject elicited properties can themselves be regarded as models of the featural organization of concepts, i.e., as models of the property spaces that shape the structure and representation of human semantic memory. The major aim of our research is to investigate the correlation between these human-derived models of property spaces and corpus-based computational models. To this purpose we have used two different sets of subject-generated properties. The first one comes from the feature norms of McRae et al. (2005), a well-known resource in cognitive science. The second set is instead represented by a corpus of image labels collected on the Web in the context of the ESP Game initiative (von Ahn & Dabbish 2004). We will now provide a brief descriptions of these property spaces, followed by a more detailed analysis of their complementary character.

2.1. NORMS: The subject elicited feature norms of McRae et al.

The semantic feature norms described in McRae et al. (2005, henceforth NORMS) are the largest set of norms available to date [1]. NORMS includes semantic features collected from approximately 725 participants for 541 living (*dog*) and nonliving (*car*) basic-level

concepts. Each normed concept corresponds to an English noun. The selection of nouns covers items most commonly used in various types of experiments on semantic memory. NORMS data were collected through a questionnaire asking subjects to list features that would describe target concepts (presented as words). The instructions also included examples of the types of properties that might be listed (e.g., physical properties, parts, etc.). Crucially, the stimuli were presented out of context, apart from homographs (e.g., *bat*), which were accompanied by a short textual clue to the relevant sense. Participants were students of various Canadian and American universities. Each concept was normed by 30 subjects.

The collected data underwent manual revision by the experimenters to normalize the subjects' productions, e.g., by ensuring that synonymous features were coded identically (e.g., *used for transportation* and *used for transport* were turned into an identical string). Features were made more explicit to ensure a better identification of the property type (e.g., *has* was added to productions describing parts of an object, such as *has legs*). In a later phase, the collected features were also classified with respect to the basic semantic type of the property. McRae et al. (2005) adapted the taxonomy of property types developed by Wu & Barsalou (in press cf. Section 4 and Appendix B for more details). NORMS also includes a number of measures characterizing the distribution of properties for the various concepts, such as feature distinctiveness (i.e., the number of concepts in which a property appears), number of distinguishing features for each concept, etc. The most relevant statistic for our analyses is the 'feature production frequency', i.e., the number of subjects out of 30 participants that listed a property. This measure is used by McRae et al. 2005 to rank the properties of each concepts, and we based on it the selection of the properties for the analysis in Section 4. As an example, Table 1 reports the top properties of the concept *car* in NORMS.

Table 1. Top 5 properties for the concept *car* in NORMS, together with their semantic types and production frequencies.

Concept	*Top properties*	*Property types*	*Production frequency*
car	used for transportation	function (*sf*)	19
	has wheels	external component (*ece*)	19
	has 4 wheels	external component (*ece*)	18
	has doors	external component (*ece*)	13
	has an engine	external component (*eci*)	13

2.2. Describing pictures: the ESP Game

ESP, the second property space we used, was built from a larger set of image descriptors collected within the ESP Game initiative (von Ahn & Dabbish 2004)[2]. The ESP Game is an attempt to label images on the Web through volunteer contribution by Internet users. The initiative is close in spirit to other enterprises (e.g., Wikipedia, Open Mind, etc.) that resort to on-line collaborative work to collect various types of knowledge. The ESP Game has however at least two peculiar features. First, its goal is to label images with words describing their content, for the long-term purpose of improving image search. Second, users label the images by playing an online game.

The game is played by two randomly matched partners that see the same image and are not allowed to communicate. Players must guess the label their partners are typing for each image. When the partners have agreed on a label, they get points and move on to the next image. They must try to agree on as many images as they can in 2.5 minutes. Players are free to use whatever word they want, except for those that belong to the list of so-called 'taboo words' for an image. This set includes those words that have already been associated to that image by other players. Taboo words guarantee a large variation in the labels associated to an image. The images presented to the players belong to a collection of 350,000 pictures randomly downloaded from Google. Images can be of all possible sorts: portraits, objects in context, landscapes, etc.

The most relevant aspect of the ESP Game is that the players are never explicitly asked to describe the image. They just have to guess what the partner is thinking and writing (hence the suggestive name 'ESP' for "extra-sensorial perception"). However, since the image is the only thing that the partners share, the most natural way for them to coordinate their minds is to type words corresponding to salient features of the image content. The evaluation by von Ahn & Dabbish (2004) indicates that, indeed, "the string on which two players agree is typically a good label for the image". The game by-product is a large corpus of images associated with all the labels the players agreed on[3]. Some examples of this output are: *speaker, hear, audio, sound, speakers, black, button* (description of music speakers); *band, guy, group, men oboe, music, hair, flute, violin, instrument, gray* (music ensemble); *eat, table, people, wine, dinner* (group of people eating).

For our purposes, the data collected through the ESP Game are a sort of *de facto* property norms. ESP labels are descriptions of salient features of the entities appearing in the images. Thus, they constitute

a model of the human property space which is elicited from human subjects in a thoroughly spontaneous and uncontrolled way.

The ESP property space we analyze comes from a random sample of ca. 363,000 labels from the whole ESP corpus. The labels are organized into 50,000 sets, each set referring to the same image. The labels in the original corpus were not lemmatized. The only processing we performed was to discard all the sets containing words such as *logo*, *ad*, *sign*, *label*, etc., since in logos and other icons an entity can be represented in a totally different way from its actual nature (e.g., a banana can be blue, etc.).

For each label pair, we count the number of distinct label sets (i.e., image descriptions) in which both labels occur. In order to downplay the importance of frequent, generic labels, we transform these raw counts into log-likelihood ratio scores measuring the association strength between two labels. Such scores are used to rank the labels associated to a given target noun. Thus, the labels associated with a target noun (a label in itself) are taken as a characterization of the properties of the corresponding concept. Table 2 reports an example of the top 5 labels associated with the noun *car* in the ESP corpus.

Table 2. Top 5 labels co-occurring with car in ESP together with their semantic types and their association strength measured by log-likelihood.

Concept	*Top properties*	*Property types*	*Log-likelihood*
car	wheel	external component (*ece*)	12.7
	road	location (*sl*)	11.4
	truck	coordinate (*cc*)	10.9
	wheels	external component (*ece*)	10.2
	race	associated event (*sev*)	9.7

2.3. Comparing NORMS and ESP

NORMS and ESP both consist of ranked lists of verbal descriptions of concept properties. Nevertheless, they differ in various respects, mostly stemming from the way these data were collected.

First of all, NORMS were elicited in an experimental situation and the subjects were explicitly instructed to generate properties for a number of concepts. *Vice versa*, the elicitation context of ESP was totally spontaneous, and the players were not told to describe the images or any features of the objects. The game task is only to coordinate with the partner. The fact that labels end up describing properties of some entity in the picture only emerges as a consequence of the

subjects' tendency to focus on salient aspects of the picture they are describing. Moreover, target images are a random sample from the Web, and thus there is no guarantee that they form a balanced set of concepts, nor that they represent prototypical instances of objects. Both the spontaneous nature of the task and the lack of control in stimulus generation make ESP more similar to corpora than to elicitation experiments.

Secondly, NORMS and ESP were obtained in two very different property generation tasks. In the former, the subjects produced the properties of a concept expressed by a noun written on the questionnaire. Conversely, in ESP the properties were produced by players observing an image, i.e., in a sort of 'implicit' picture description task.

Last but not least, the property sets in NORMS were elicited by presenting the concept nouns 'out of context' (apart from few cases of homography). Conversely, most of the pictures labeled in ESP represent 'situated entities', i.e., entities with a context, such as for instance a cow in a meadow, a person driving a motorbike or drinking beer, etc. In some cases, there is a figure clearly emerging from the background, while other pictures simply contain a large scene with different entities involved in some activity. Since no instruction is provided about which entity is to be described, the players are free to parse complex scenes as they please, and focus on specific objects with the only constraint of maximizing the probability to converge on the partner's choice.

The differences between NORMS and ESP are particularly relevant in the light of the recent debate in cognitive science on the 'situated' nature of conceptualization (Glenberg & Kaschak 2002; Barsalou 2005; Wu & Barsalou in press). According to the situated cognition view, concepts are grounded to some extent on sensory-motor systems, and properties rather than being abstract amodal symbols are themselves grounded in perception and action. Wu and Barsalou (in press) bring behavioral evidence showing the strong correlation between properties generated by subjects explicitly instructed to use mental images and the properties produced by subjects that did not receive such an instruction. These results are interpreted as supporting the view that subjects generate properties of a concept by "running" perceptual simulations of its instances. Moreover, Wu and Barsalou show that an average of 25% of the properties produced by their subjects are related to aspects of the prototypical contextual setting of the concept instances, such as typical actions and locations, entities co-occurring in the same context, etc. This fact is taken by the two authors as evidence that "Rather than being decontextualized and stable, conceptual representations are contextualized dynami-

cally to support diverse courses of goal pursuit" (Barsalou 2005: 622).

In the next sections, we will tackle the issue of how computational property spaces correlate with subject-generated semantic feature sets. However, our analysis will also focus on the comparative analysis of the types of properties in ESP and in NORMS. In fact, the peculiar characters of these two models suggest that their comparison can provide interesting evidence on the relationship between conceptual representations and perceptual features (notice how ESP is by design a strongly "situated" property space), as well as on the interplay between concepts and context.

3. Word space models

Corpus-based "word space models" (Sahlgren 2006) induce the semantic representation of words from their patterns of co-occurrence in text. The meaning of a word is thus represented by a vector whose dimensions are co-occurrence scores or a function of co-occurrence scores. Standard geometrical methods can then be used to assess semantic similarity in the vector space.

It is worth pointing out that cognitive work has concentrated on 'concepts', rather than 'word meaning', that is instead the focus of word space models. However, the two notions are close enough (see discussion in Murphy 2002) that we will apply standard word space models to what cognitive scientists might see as a "conceptual" task. The issues with a direct comparison of properties generated by humans and computational models discussed in Section 4.1 below largely arise from differences in the way in which conceptual properties can be lexicalized, and an important problem we gloss over here is that words tend to be polysemous, and thus point to sets of concepts rather than single concepts.

We work with two word space models representing different traditions. The model we call SVD takes a window-based view of co-occurrence, where any word that occurs within a certain distance to the left or right of the target is treated as context. Since this will typically lead to a very large and sparse co-occurrence matrix, models of this sort benefit from dimensionality reduction techniques such as singular value decomposition.

The StruDEL model takes instead a pattern-based view of co-occurrence, treating as potential contexts only those words that are connected to the target by patterns that might cue an interesting semantic relationship. While general word space models rarely adopt

this approach (we are only aware of the line of research summarized in Poesio & Almuhareb 2008), pattern-based methods are common in studies that attempt to identify 'specific' types of semantic relations, at least since the seminal work of Hearst (1992) on the hyponymy relation.

There is a large number of alternative word space models. We are not claiming that the ones we selected are the best or most interesting ones. However, we do believe that they are fairly representative of the two approaches we just sketched, that, in turn, account, with important variations, for most models we are familiar with.

In particular, the 'Hyperspace Analogue to Language' (HAL) model (Burgess & Lund 1997) is similar to our SVD, without dimensionality reduction (but with dimension weighting). The popular 'Latent Semantic Analysis' (LSA) model (Landauer & Dumais 1997) is similar to our SVD, except that co-occurrence is measured in terms of documents rather than word windows. Window-based dimensionality-reduced models have been shown to outperform both non-reduced and document-based models at least in the classic TOEFL synonym task (Rapp 2003, 2004).

In the dependency-based model of Pado & Lapata (2007) only words that are linked by specific syntactic relations are treated as potential contexts. This model is intermediate between the window-based approach, that is purely based on syntagmatic linear order, and the pattern-based approach, that tries to zero in on semantically meaningful contexts.

3.1. SVD

Our SVD model is based on a lemmatized version of the BNC [4] with only content words (nouns, verbs, adjectives) preserved. The 21,000 most frequent words in this version of the corpus (minus the top 10 most frequent words) are treated as targets, i.e., words for which we build a semantic representation. The top 2,000 words (minus the top 10 most frequent ones) are treated as potential contexts, i.e., the words whose co-occurrence with the targets is recorded.

We build a target-by-context co-occurrence frequency matrix, counting only instances in which a potential context word occurs within a window of 5 words from a target. The co-occurrence matrix generated in this way is then reduced using singular value decomposition. The reduced matrix has 21,000 rows (the target words) and

125 dimensions (the 125 left singular vectors that account for most of the variance, multiplied by the corresponding singular values). The word space is constructed using the Infomap tool [5].

In a previous experiment with the widely used TOEFL synonym set, the same SVD model we are using here reached accuracy around 91.3%, comparable to the best performance on this task reported by Rapp (2003). Thus, we are experimenting with a state-of-the-art SVD-based model.

What are the 'properties' of concepts in SVD? The most straightforward approach would be to treat the reduced space dimensions as properties. However, these dimensions are hard to interpret. An attempt in this direction would be to look at the *n* words that have the highest and lowest values on a dimension, to get the gist of what the dimension is about. A preliminary analysis along these lines of the top 10 dimensions and of a random sample of 10 other lower ranked dimensions suggests that this approach will not work for our purposes. This becomes clear by looking at Table 3, that reports the top (positive valued) and bottom (negative valued) 5 words associated to (randomly chosen) dimensions 5 and 15.

Table 3. Words with 5 top and lowest values on dimensions 5 and 15 of SVD model.

Dimension	*Top words*	*Bottom words*
5	political, rhetoric, ideology, thinking, religious	around, average, approximately compare, increase
15	juice, colouring, dish, cream, salad	police, policeman, road, drive, stop

Table 3 clearly illustrates two problems with treating dimensions as properties. First, they correspond to broad domains or topics (intellectual life, quantities, food, car traffic…) rather than to specific properties (the classification by domain is orthogonal to the one by property type). Second, each dimension tends to do double duty (at least), with positive value locked onto one domain and lower values locked onto another, unrelated domain (it is hard to see a relation between, say, food preparation and traffic) – conversely, although it is not illustrated here, we found several cases in which different dimensions pointed to the same domain. These findings essentially confirm the fairly common statement in the literature that the dimensions of SVD matrices are not directly interpretable as semantic features (Kintsch 2001). Instead, the only viable way to explore the meaning of a vector is by inspecting the words that appear close to it in the semantic space.

Therefore, we took the nearest neighbours of a word in the Euclidean space defined by the dimensions (with cosine as the nearness measure) to be the SVD-produced 'properties' of a word/concept. Property identification is not one of the tasks that word space models of this sort were designed for, and we realize that their proponents could argue that we are putting them to an improper usage. However, to the extent that properties are an important aspect of concepts, the nearest-neighbour-as-property approach is the most natural one for SVD and related models.

Continuing with the *car* example, the top 5 properties of this concept in the SVD space are listed in Table 4.

Table 4. Top 5 properties (=neighbours) for the *car* concept in SVD, together with their semantic type and cosine.

Concept	*Top properties*	*Property types*	*Cosine*
car	van	coordinate (*cc*)	.75
	driver	participant (*sp*)	.73
	vehicle	superordinate (*ch*)	.71
	park	action (*sa*)	.70
	motorist	participant (*sp*)	.69

3.2. StruDEL

Whereas SVD is a 'garden variety' word space model, of the sort often encountered in the literature, the StruDEL model (for Structured Dimension Extraction and Labeling) is first proposed here. We will not argue for the virtues of StruDEL (it does have many, but they will be presented elsewhere), but rather use it as the representative of an approach to word space models that differs from the 'flat co-occurrence' of SVD, being based on the search for semantically meaningful patterns. As we already mentioned, StruDEL should be seen as a generalization of the pattern-based approach to information mining used by Hearst (1992) and many others.

StruDEL builds structured word spaces in two phases. First, it uses pattern matching to find and rank potential properties of the target words (concepts). Then, it generalizes from the strings connecting concepts and properties to find (lexical correlates of) the relation that links them. One fundamental intuition behind StruDEL is that true semantic relations will be expressed by a variety of surface realizations. Thus, rather than ranking properties on the basis of token frequency, it ranks them on the basis of the number of distinct patterns that connect them to the target concepts.

Given a list of target nouns and a (POS-tagged) corpus, StruDEL looks for nouns, adjectives and verbs that occur in the near of a target. Only words that are linked to the target by a 'connector pattern' that follows one of a limited set of templates are considered potential properties.

The templates for nominal properties are simple regular expressions that specify that the target and property must either be adjacent (the noun-noun compound case) or they must be connected by a (possibly complex) preposition, or a verb, or the possessive (*'s*), or a relative such as *whose*. Optional material, such as adjectives and articles, can occur in the connector pattern, whereas other categories, such as names and sentence boundaries, act as a barrier blocking the potential template match. The template matching component also performs basic pattern normalization by replacing all verbs and adjectives that are not in a 'keep list' of 50 frequent verbs and 10 frequent adjectives with the corresponding POS tags. Table 5 presents (somewhat simplified)[6] examples of the extraction procedure for the concept *onion* and the candidate property *layer*. Similar rules are applied to the extraction of adjective and verb properties.

Table 5. Examples of input and output to the StruDEL pattern template component. Notice the 'Position' field, included in the pattern and recording whether the concept is the word to the left (*onion with different layers*) or right (*layer from an onion*).

Input	*Output*		*Notes*
	Pattern	*Position*	
layer from an onion	from a	right	*an* normalized to *a*
layers in a red onion	in a JJ	right	*red* mapped to *JJ*
onion with different layers	with different	left	frequent adj *different* preserved
onions and with their layers	Ø		conjunction blocks pattern extraction

In the next and crucial step, concept-property pairs are ranked based on the number of distinct patterns that link them, ignoring the token frequency of the concept-property-pattern tuple. The intuition behind this approach is that a single, frequent concept-pattern-property tuple could simply be a fixed expression, or more in general a combination that is frequent for accidental reasons. On the other hand, if concept and property appear with many distinct patterns, i.e., their relation is predicated in many different ways, it is more likely that they are connected by an inherent semantic link. For example, *year of the tiger* is much more frequent in our corpus than any pattern

connecting *tail* and *tiger*. However, *year of the tiger*, because of its idiosyncratic nature and proper-noun-like usage, is the only attested pattern linking these two words (we do not find: *year of some tigers*, *tigers have years*, etc.). The relationship of tigers with tails, instead, is expressed in a number of ways: *tail of the tiger*, *tail of a tiger*, *tigers have tails*, *tigers with tails*, etc. Pattern type frequency is a better cue to semantics than token frequency.

More precisely, our rank is based on the strength of the statistical association between concepts and properties sampled from the list of distinct tuples (akin to sampling concepts and properties from a dictionary of distinct longer strings rather than from a corpus). Association, measured by the log-likelihood ratio statistic, is better than raw frequency since it weights down properties that might occur in a number of patterns simply in virtue of their generic nature (e.g., *year* and *time*, that can occur with almost anything). For practical reasons, we preserve only those properties that are very significantly ($p < .00001$) associated with a concept.

In the next step of the StruDEL procedure, we provide a shallow description of the relation occurring between a concept and a property by generalizing across similar patterns that connect them, and keeping track of the distribution of these generalized patterns in what we call the 'type sketch' of the pair (the generalized patterns are seen as shallow cues to relation 'types'). We are following here a long tradition in lexical semantics proposing that semantic relations can be captured directly by the explicit syntactic material expressing them (see, most notably, Levi 1978). We store the whole type distribution associated with a concept-property pair, rather than the most common type, because this is useful for disambiguation purposes (*in* might cue hypernymy in a sketch with *such_as*, but location if it occurs with *on)*.

Generalization is performed by another simple rule-based module that essentially looks for prepositions, verbs and other 'meaningful' components of a pattern. Consider a hypothetical concept-property pair occurring with the following patterns: *with a number of* (2 times), *with a* (1 time), *with JJ* (1 time), *have* (1 time) and *has* (1 time). The type sketch for this pair would be: *with* (66.6%), *have* (33.3%). Illustrative examples of the StruDEL output, including type sketches, are presented in Table 6.

Table 6. Type sketches: properties are annotated with part of speech; log-likelihood is the concept-property association score computed as described in the text; types are suffixed with position of concept in relation, and only types accounting for at least 10% of the distribution are presented.

Concepts	*Properties*	*Log-likelihood*	*Type sketches*
child	parent-n	11726.7	of+right (40%), with+right (11%)
child	parent-v	120.8	_+right (79%)
lion	mane-n	259.1	's+left (50%), with+left (15%), have+left (12%), of+right (10%)
egg	female-n	1603.4	produce+right (13%), by+left (12%)
breakfast	croissant-n	257.2	for+right (46%), of+left (34%), with+left (12%)
beach	walk-v	687.6	_+right (29%), from+right (24%), along+right (23%), on+right (13%)
grass	green-j	277.6	_+right (58%), is+left (25%), is_ADV+left (16%)

Thanks to type sketches, StruDEL can be tuned to different semantic tasks (e.g., in a telic quale task, one could pick only properties with *for* as a prominent type in the sketch). However, here we just use them as a filtering device: We weed out from the model those concept-property pairs whose dominant type in the sketch is not among the top 10 most common types in the whole StruDEL output list.

We created a StruDEL semantic space using the 542 concepts of McRae et al. (2005) as targets. Model statistics were extracted from the large, Web-derived ukWaC corpus (about 2.25 million tokens)[7]. Notice that in a series of preliminary clustering experiments we also trained the SVD model on these data. However, ukWaC-based SVD performed systematically worse than BNC-based SVD (StruDEL's pattern extraction component probably acts as a 'junk filter', that makes this model more robust to the noise inherent to Web data, whereas SVD, taking any context into account, is not as robust).

Given that StruDEL is explicitly designed to represent concepts in terms of their properties, the evaluation conducted here is entirely straightforward: we pick and analyze the top 10 properties (ranked by log-likelihood ratio and filtered by common type as described above) of each target concept.

The top 5 properties of *car* for StruDEL are presented (without type sketches) in Table 7.

Table 7. Top 5 properties for the concept *car* in StruDEL, together with their semantic types and association strength measured by log-likelihood.

Concept	*Top properties*	*Property types*	*Log-likelihood*
car	drive	activity (*sa*)	1795.4
	driver	participant (*sp*)	1329.7
	park	activity (*sa*)	985.4
	road	location (*sl*)	839.3
	garage	location (*sl*)	704.3

4. Property Analysis

4.1. Design and materials

We selected 44 concrete nouns belonging to 6 semantic categories from the feature norms in McRae et al. (2005): 4 categories of natural entities (*birds*, *ground animals*, *fruits*[8] and *greens*) and 2 categories of artefactual entities (*vehicles* and *tools*). We assigned the nouns to their category, since no classification was available in the norms. The complete list is reported in Appendix A. The mean frequency of the nouns in the BNC is 3,320 (σ = 5,814). The noun with the lowest frequency is *chisel* (233) and the one with highest frequency is *car* (35,374). ANOVA revealed no significant difference between the six semantic categories with respect to the log-frequency of their elements (F = 1.0964, p = 0.3784). We then extracted the top 10 properties associated with each noun in NORMS, ESP, SVD and StruDEL, obtaining 1,727 distinct concept-property pairs (some pairs are repeated across spaces, and some concepts are associated with less than 10 properties in ESP).

Analyzing the specific properties associated to the concepts would seem the most straightforward way to compare the property spaces. However, this solution proved not to be viable in practice. In fact, in preliminary experiments with direct properties, the overlap among human and computational models was never above 21%, and the correlation among ranks of overlapping properties was not above 0.16[9]. These low values are partially due to genuine differences among spaces, but they are also often due to normalization problems. For example, if one space lists *noisy* as a salient property of helicopters, whereas another space includes *loud*, it is extremely hard to determine by automated means that these are different lexicalizations of the same property. Moreover, an analysis at such a granular level would not

allow us to see the generalizations in the kinds of properties that different spaces assign to different concept categories.

These considerations prompted us to compare the property spaces at a more abstract level, i.e., at the level of 'semantic types'. Therefore, the properties extracted from ESP, SVD and StruDEL were classified according to the hierarchical coding scheme used in McRae et al. (2005). For NORMS, we simply adopted the classification available in McRae et al. (2005) (cf. Section 2). The classification (reported in Appendix B) consists of an 'ontology' of property types organized under 4 main classes:

- *category (c)* – properties providing taxonomic information about a target concept (e.g., its superordinate concept);
- *entity (e)* – properties describing an entity's internal and external composition, typical behaviour, etc.;
- *situation (s)* – properties referring to aspects of the contextual situation in which an entity may appear (e.g., typical function, other entities co-occurring in the same scene, actions performed on an entity, typical location, etc.);
- *introspective (i)* – properties describing a subject's mental or affectional state towards an entity.

A special category *Out* has been added to the original scheme, to mark those cases in which the property is not prototypically related to the target concept. Obviously, this class never occurs with the properties extracted from NORMS. Conversely, *Out* cases are variously attested in the other property spaces, mainly as a consequence of the computational processes used to generate them.

The concept-property pairs from different spaces were merged before annotation, to avoid biases coming from our *a priori* expectations about the models. Moreover, to minimize differences between the annotation of McRae and colleagues and ours, we adopted their labels for pairs in their database and present in other spaces as well and, more in general, we looked at the choices made in their database as our main source of guidance and annotation policies.

We independently annotated each concept-property pair, and discussed all the cases of disagreement. After a few rounds of training in applying the classification scheme to random samples extracted from the concept-property pair set, we decided to merge synonym, coordinate and subordinate properties under the common type *coordinate* (*cc*). This change was prompted by the complexity of discriminating between these fine-grained property types out of context (is *tiger* a hyponym or a co-hyponym of *cat*?), potentially resulting in coding inconsistencies.

Of course, several classification decisions were rather difficult. Often these difficult choices cut across the main classes of the ontology. For example, are bowls and pans coordinates (*cc*) or situationally associated entities (*se*)? Is cutting the function of scissors (*sf*) or their typical action ('behaviour': *eb*)? Unfortunately, the ontology misses natural classes cutting across the proposed major categories, such as 'related entities' – that can be taxonomically related (*cc*) or situationally related (*se*) – and "activities" (*eb*, *sa*, *sf*). Our analysis will focus more on such natural classes – that let important generalizations about the nature of different spaces emerge – than on the high level categories of the ontology shown above.

4.2. Data cleaning

Out cases were unequally distributed, accounting for 30% of the properties in ESP, 11% in SVD and 7% in StruDEL. We attribute the over-representation of *Out* in ESP to the fact that often ESP pictures describe complex scenes. For example, *sky* comes up as one of the top properties of elephants since they are more likely to be photographed outdoors. Having ascertained this, we looked at whether the distribution of *Out* cases across categories (ground animal, fruit, etc.) changed from property space to property space. A logistic regression with concept category, property space and their interaction as factors and *Out* responses as independent variable showed that *Out* is significantly ($p < 0.01$) more likely in ESP than in SVD or StruDEL, and that tools are significantly ($p < 0.05$) more likely to trigger *Out* responses than the other concept types (probably because they occur almost by definition in complex scenes). Importantly, there is no significant interaction. Thus, we can remove the *Out* cases from the analysis without inserting a bias in the model comparison.

In order to avoid sparseness problems and to simplify the analysis, we decided to ignore rarely used property types. Choosing a cut-off point was easy, since we observed a large interval between property types *eb*, that occurs 85 times in total across the spaces, and *sp*, that occurs 36 times only. Thus, we removed the latter and all the rarer properties, i.e.: *eae, eci, em, eq, esi, esys, ew, ie, in, io, sev, sp, st* (refer to Appendix B for the codes). The full frequency table, including the rare types, is presented in Appendix C.

4.3. General distribution of properties

We first look at the overall distribution of property types across

property spaces. Table 8 reports X^2 values computed on pairwise space-by-property contingency tables. The smaller the value, the better the fit between two spaces in terms of property type distribution (none of the fits is particularly good in absolute X^2 terms, but we are interested in relative comparisons).

Table 8. Pairwise X^2 fits among spaces.

	NORMS	ESP	SVD	StruDEL
NORMS	–	144	431	208
ESP	144	–	140	143
SVD	431	140	–	261
StruDEL	208	143	261	–

The first interesting datum is that ESP is (comparatively) close to each of the other spaces. As we will see below, ESP looks like a sort of 'average' model with no single property type that is seriously over- or under-represented with respect to the other spaces. To the extent that we think that all other spaces have something going for them, this makes ESP rather attractive as a 'balanced' space (keep in mind, however, that we are analyzing a 'cleaned' version of the ESP space, that would otherwise be characterized by about 1/3 *Out* cases: see 4.2 above). ESP is similar to NORMS in that it is based on human-elicited data; however, ESP concept-by-property characterizations are implicit in patterns of co-occurrence of words in descriptions of random images and have to be extracted with distributional techniques similar to those used for corpora. This double nature gives ESP an intermediate status among property spaces. Interestingly, ESP is closer to both SVD and StruDEL than the two corpus-based models are to each other.

Strikingly, StruDEL has a better fit to NORMS than to SVD, the latter being the 'outlier' space, the one most distant from both NORMS and StruDEL. We have here an argument for StruDEL as a better approximation to the human property space. This is not surprising, given that StruDEL, unlike SVD, was designed to capture properties. More importantly, this result warns against treating 'corpus-based' models as a monolithic whole, assuming that, no matter how much they differ, these differences will not be as large as those between humans and distributional models. The data in Table 8 show clearly that this is not the case. Any conclusion one might reach about a specific corpus-based model will not necessarily apply to other corpus-based approaches as well.

We take now a closer look at the property types that characterize each space using the summary in Figure 1. This is a 'mosaic

plot' (Meyer et al. 2006) that visualizes the property-space-by-property-type contingency table through rectangles whose areas are proportional to observed frequencies. Each row represents a property space. The columns correspond to property types, with type labels at the top of the plot and redundantly coded inside cells that are large enough to allow this (if the cell is too narrow, its property type must be inferred from the list at the top of the plot and/or by the labels of the surrounding cells: for example, the second rectangle of the SVD row represents the SVD-by-*ch* count). Grey shadings are used to highlight strongly over- or under-represented cells (Zeileis et al. 2005); in particular, cells with absolute Pearson residuals (quantifying the contribution of a single cell to the X^2 statistic) between 2 and 4 are light grey, and cells with Pearson residuals above 4 are dark grey (Pearson residuals approximate a standard normal distribution, thus the 2 and 4 thresholds correspond, approximately, to 0.05 and 0.0001 significance levels).

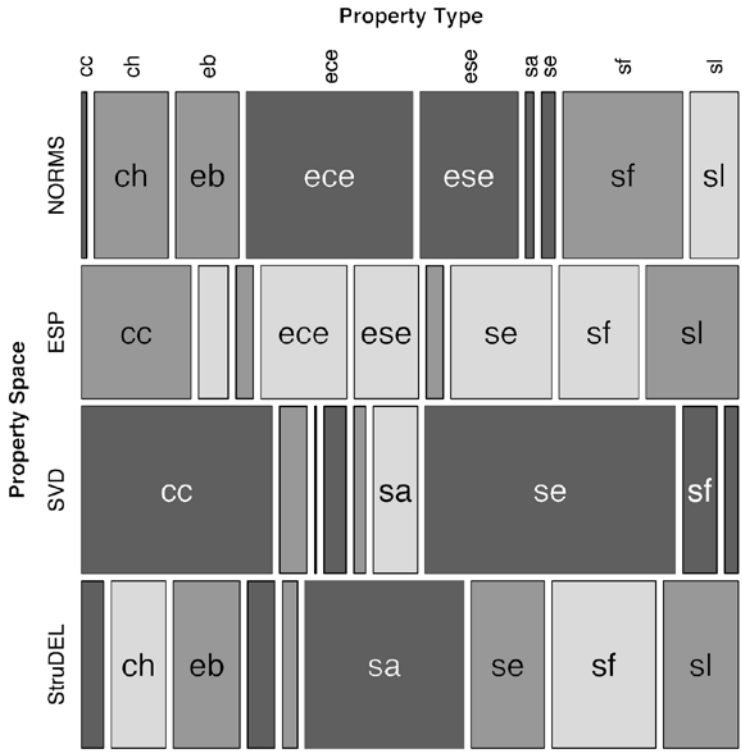

Figure 1. Distribution of property types across property spaces.

Looking at NORMS first, we notice the relatively high frequency of external components (*ece*) and surface properties (*ese*), and the almost complete lack of coordinate (*cc*) and situationally related (*se*) entities. External parts (like the wheel of a car) and surface properties (like the fact that a banana is sweet and yellow) are obviously perceptually important characteristics of concrete concepts, and they are almost completely missed by our corpus-based models. This suggests an important line of research in improving such models, perhaps incorporating visual cues into the distributional statistics (the ESP space does not have a similar problem). Coordinate and situationally related entities, on the other hand, might be triggered by free association tasks (*dog* in response to *cat*) but they are unlikely properties in a concept description (*dog* as a characteristic property of *cat*). In this case, the problem is mainly with the SVD space, where *cc* and *se* are by far the most common property types. Interestingly, in this respect StruDEL is the closest model to NORMS (having a lower number of coordinate and related entities), whereas ESP does have its fair share of these properties (not surprisingly, given that pictures are scene descriptions and scenes are quite likely to include coordinate – e.g., different kinds animals – and related entities – e.g., spoons and bowls). Another similarity between NORMS and StruDEL pertains to the hypernym (*ch*) and entity behaviour (*eb*) properties, that are well attested in these models only. Both functions (*sf*) and locations (*sl*) are well represented in NORMS as well as ESP and StruDEL, whereas SVD under-represents both.

It is intriguing that situated categories (*sa, se, sf, sl*) account for about one fourth of the NORMS properties, in a very good match with what has been reported by Wu & Barsalou (in press). All other spaces have a higher proportion of situated properties. Still, given the considerations we made in Section 4.1 on the spurious nature of high level categories such as *situation*, we are not sure of how meaningful this observation really is. For example, SVD features mostly *se*'s, that are arguably closer to 'categorical' property *cc* than to situational property *sf* – function – that is instead more typical of ESP and StruDEL.

Turning now to ESP, the plot confirms that this is in many respect the "average" space, with no cells that deviate from the expected values in a highly significant way. The most common types are those that are useful to describe an object in a photographic context: coordinate and related entities (*cc* and *se*), external parts and properties (*ece*, *ese*), location (*sl*). Interestingly, the function of objects (*sf*) is also well represented (perhaps scenes captured in pictures tend to show objects engaged in their characteristic function: e.g., bottles

are more likely to appear in scenes where somebody is drinking?). Other 'activity' properties (*eb* and *sa*), on the other hand, are almost absent. This might be due to the fact that entity behaviours and activities will often not be captured by static pictures, although this claim deserves further investigation (it is not hard to imagine, say, pictures of somebody training a dog or driving a car). ESP is also under-representing hypernyms (*ch*), which might be explained by the contingent and visual nature of salient information in pictures (one picture might be described as a brown dog wearing socks, but the fact that the dog is an animal would not add much to the picture description).

Coming to the computational models, as we said, SVD is clearly the outlier here, with most of its properties being related entities – either taxonomically related (*cc*: dog and cat) or situationally related (*se*: spoon and bowl). This is not surprising given that SVD does so well in synonym detection tasks (synonyms are the limiting case of taxonomic relatedness). SVD neighbours (that we are interpreting as properties) are based on paradigmatic similarity, i.e., the tendency to occur in the same linguistic contexts. However, a concept (say, *dog*) and its properties (say, *tail*) rarely occur in exactly the same narrow context: they will rather occur *near* each other. In future studies, we would like to check whether a SVD model based on larger context windows (dozens of words, or whole documents), and thus exploiting information from a wider syntagmatic span, can capture other kinds of properties beyond related entities.

StruDEL presents a more balanced property space, with counts comparable to those in NORMS for hypernyms (*ch*), typical behaviours (*eb*), function (*sf*) and locations (*sl*). Like NORMS, StruDEL avoids coordinates (*cc*) and, to a lesser extent, situationally related entities (*se*). On the other hand, StruDEL misses external components and properties (*ece*, *ese*) almost completely. An interesting feature of this model is how it is the one that highlights functional/activity properties the most: *eb*, *sa*, *sf*. This is probably due to the fact that StruDEL includes rules that specialize in the extraction of properties expressed by verbs from corpora.

To conclude, the previous analysis has taken NORMS as the *de facto* gold standard on which to evaluate the other spaces. However, the fact that different spaces specialize in different kinds of properties (and, with the exception of ESP – that has a sizable *Out* problem – they produce relevant kinds of properties) is in other respects very positive. In particular, we can use the models/spaces in a complementary way in order to build a rich and multi-faceted view of human semantic cognition.

4.4. Properties of different categories of concepts

The distribution of property types across concept classes can be best observed in Figure 2, that contains "association plots" between property types and concept categories for each property space. Association plots (Meyer et al. 2003) represent the pattern of deviation from independence between two categorical variables by visualizing the table of Pearson residuals derived from their contingency table, thus showing the net contribution of each cell to the global X^2 statistic. Rectangles have height proportional to the corresponding Pearson residual. The sign of the residual – i.e., whether a cell is over- or under-populated – is coded by the position relative to the baseline. Width is proportional to the square root of the expected frequencies, so that the areas of the rectangles are proportional to the non-normalized difference between observed and expected values. The most important information for our purposes is encoded in the heights, that represent, for each space, the degree to which the observed frequencies of property types for a concept category depart from the counts expected given the overall distribution of the space (i.e., departures from the space-specific distributions depicted in the rows of Figure 1). Like in Figure 1, the shades of grey of the rectangles correspond to Pearson residuals with absolute values larger than 2 (light grey) and 4 (dark grey) respectively. We will concentrate our analysis on such large residuals, cueing the most salient properties of each concept category.

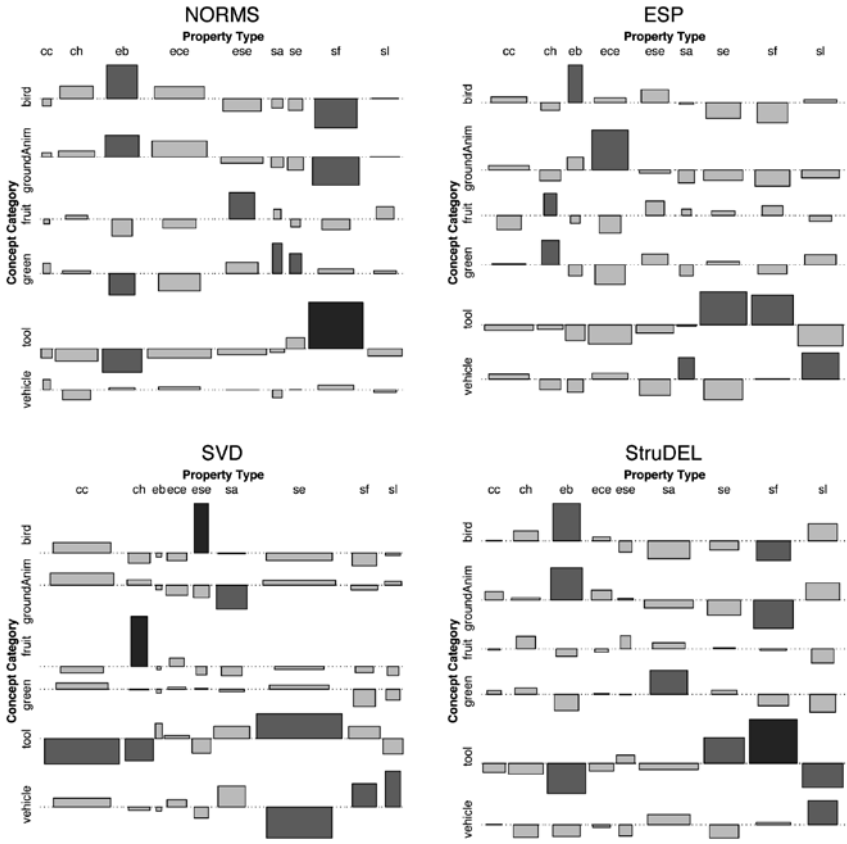

Figure 2. Positive and negative deviations from the expected value of properties across concept categories, for the four models.

The association plot for NORMS shows a major contrast between animals and tools, with the latter being mainly characterized by functional properties (*sf*), that are instead strongly under-represented in birds and ground animals. Zooming in on the natural domain, animals are more reliant on properties referring to their typical behavior (*eb*), which are instead below expected distribution with vegetable categories. The latter concepts do not behave homogeneously. Fruits are mostly characterized by properties referring to their external surface, while vegetables show a significant dominance of properties referring to typical activities (typical ways of cooking and eating them, e.g., boiling for potatoes, etc.) and to associated entities. Interestingly, vehicles are the only category that is not identified by any specific

property type (except for a weak presence of *eb* properties, such as flying for helicopter).

We can compare the distribution of property types by concepts classes in NORMS with the data reported in Vinson et al. (2003), resulting from the analysis of semantic feature norms elicited for a set of objects and action concepts. Vinson et al. (2003) claim that artifacts have significant more functional features than natural concepts. Although they use a different property classification scheme, their results are strikingly similar to NORMS with respect to tools and its mirror image represented by animals. Vinson et al. (2003) also observe that visual features (i.e., referring to the sense of vision) are more significantly associated with animals, fruits and vegetables. These results are partially confirmed in NORMS. The *ese* properties that dominate fruits actually refer to external, visible features. On the other hand, we have said above that animals are instead characterized by behavioral properties. McRae et al. (2005) cross-classified the properties in their norms with an orthogonal taxonomy referring to the brain area where properties are plausibly computed. Interestingly, in this parallel classification scheme most of the *eb* properties appearing with our animal concepts are marked as visual.

The preferred association of functional properties with tools appears also in ESP, together with their symmetric under-representation in animals. However, now *sf* is not the only hallmark for tools, which are strongly characterized by associated entities too (typically, objects co-occurring with the target, e.g., knife with spoon). *Prima facie*, this is not surprising, since surely tools often appear in pictures together with other objects (take for instance pen and paper). However, this can not be the only explanation, since, in ESP images, entities belonging to any concept category appear in scenes with other objects. Yet, it seems that players single out associated entities particularly with tools (where there might be a stronger functional link between the concept and the associated entities).

An interesting parallelism between ESP and NORMS is the salience of *eb* with animals, although now association is limited to birds, while ground animals show a preference for being described in terms of their parts. Fruits and greens present patterns that are much different from to the ones in NORMS. Both classes are characterized by the over-representation of their hypernyms (*ch*), which do not play a significant role in distinguishing any category in NORMS. Moreover, now vehicles appear to be strongly characterized by location features (e.g., road for cars) and by typical activities (e.g., driving). Overall, ESP shows very distinct distributions of properties across concept

categories. This is particularly interesting in the light of the ways in which ESP labels are generated. In the ESP Game, labels are the result of an autonomous and spontaneous parsing of the scene carried out by two players. In scanning the scenes in the pictures, it seems that concept categories make different property types be more salient for players that must converge on a choice.

Coming to the corpus-based property spaces, we first notice a striking parallelism between StruDEL and NORMS, especially with respect to the key role of functional features with tools and its symmetric under-representation with animals, which instead are again globally characterized by behavioral properties. This is an important parallelism, also because the animal/tool distinction is one for which there is some of the most robust neuro-imaging evidence (Martin 2007). Other two elements of similarities concern the fact that taxonomic properties do not play a significant role with any category, and the association of vegetable concepts with typical activities. Differently from NORMS, *se* plays an important role in the StruDEL representation of tools, together with *sf*, and in close parallelism with ESP. StruDEL is also similar to the latter with respect to the prominence of location properties with vehicles. A major contrast between StruDEL and the two human-generated property spaces is instead given by the fact that no category is specifically associated with either surface properties or parts, surely a consequence of the fact that these types of properties are very rarely captured by this model.

Similarly to what we noticed in Section 4.3, the property-by-category distribution also reveals that StruDEL is more strongly correlated with the human-property spaces than with SVD, which is confirmed as a sort of outlier under many respects. For instance, SVD is the only model to characterize birds with surface properties, and the robust association of hyponyms with fruits we observe in ESP is even stronger in SVD. Conversely, this model goes together with ESP and StruDEL in having locations as significantly related to vehicles. The rather 'excentric' character of the SVD property space is however best revealed by the fact that, while the other models (although to different extents) assign a prominent role to behavioral properties for animals and to functional ones for tools, neither of these correlations is observed in SVD. Tools are now strongly characterized only by associated entities, while *sf* plays a significant role with vehicles. We already observed in the mosaic plot of Figure 1 the general bias of SVD towards paradigmatic associations, to which *se* properties belong (together with *cc*). The interesting fact now is that this bias towards *se* is not equally distributed among the various concept categories.

Like ESP, SVD tends to single out associated entities especially with tools. This points towards a more sophisticated interplay between the general tendency of SVD to highlight paradigmatic associations, and the specific semantic organization of particular conceptual categories.

To conclude, once more we found that both similarities and divergences cut across the human-/corpus-generated divide. This points, again, to a certain complementarity between models, where it is not always necessarily the case that NORMS is the 'best' one. For example, ESP and StruDEL, very sensibly, assign typical location as a salient property of vehicles, whereas this is completely missed by NORMS.

5. General Discussion

It is now time to go back to the main issue that we raised at the outset of our work, i.e., to what extent the linguistic expressions co-occurring with a word are correlated with the properties that compose its concept. The results of our analyses confirm that there is no easy answer to such a question. The reasons depend both on the cognitive construct of property as emerging from human-elicited data, and on the behavior of the computational models used to approximate such a notion. The comparison between NORMS and ESP has revealed important parallelisms, but also many equally salient divergences, both at the level of global distribution of property types, as well as with respect to specific concept categories. This fact suggests important differences within human property spaces, and warns against taking a single or specific set of human generated data as 'the' gold standard with which distributional models can be compared. A viable hypothesis is that there is actually no unique human property space, but rather a representational core that is variously modulated depending on the task, context, medium and mode of expression, etc.

Similar caveats also extend to corpus-based, distributional models. In our experiments, two different approaches to carve semantic knowledge out of corpus-based word distributions have been shown to produce highly different semantic spaces, with little if anything in common. These are not simply alternative ways to acquire semantic information from texts, but they are rather methods that extract different portions and aspects of the semantic space. StruDEL has a better fit to NORMS, but the picture becomes more complex once we also take into consideration the variation within human models,

since ESP appears to be equally distant from both computational spaces. Great caution should be used when interpreting corpus-based models as simulations of human semantic space. Since corpus-based models do not behave uniformly as far as the shape of the semantic space they produce, the specific way in which they processes corpus data and derived semantic information must be taken into account. A poor match between corpus-based data and subjects' elicitations may be the result of a complex array of factors concerning the peculiarities of the computationally (and experimentally) derived semantic space.

In Section 2.3, we reported the result from Wu & Barsalou (in press) that property spaces generated by subjects in neutral conditions are more strongly correlated to property spaces generated by subjects instructed to use mental images than to property spaces generated by subjects instructed to use word associations. Wu and Barsalou take this as evidence supporting the hypothesis that the human property space is not "amodal" – like the one based on simple word associations – but it is instead inherently grounded on perceptual modalities, through the system of perceptual simulations of concept instances that are run by subjects in producing properties. The results of our experiments warn against drawing similar straightforward conclusions. First, we saw in Table 8 that ESP, a strongly 'perceptually grounded' space, given that labels are generated by describing pictures, is equally close to NORMS as it is to SVD and StruDEL, that have all the hallmarks of 'amodal' property spaces, based purely on corpus-derived word associations. Moreover, the analysis of Figure 1 and, especially, Figure 2 suggests that there are important similarities between NORMS and StruDEL. This raises the question of to what extent property sets generated by subjects are determined by statistically significant correlations between linguistic structures to which subjects are exposed to in their communicative tasks. Our experiments do not allow us to advance any further hypothesis without the risk of being purely speculative. Yet, they suffice to highlight the complexity of the relationship between computational linguistic and cognitive research, confirming at the same time all the potential offered by their encounter.

Having conducted the qualitative analysis we presented here, we would like, in future work, to see how the different natures of the models lead to different performance in tasks such as unsupervised clustering by concept category, modeling free association or synonym detection. Given the almost complementary nature of SVD and StruDEL, for example, we would not be surprised to find out that

they succeed in modeling different aspects of semantic cognition. At a more technical level, we would like to experiment with a SVD model derived with larger context windows, to give more weight to syntagmatic, as opposed to paradigmatic, neighbours. Finally, our analysis here has ignored possible effects due to specific concept categories (say, cherries or cars) and property types (say, red or barking). We are currently exploring the possibility of using multi-level modeling techniques for an analysis that takes these effects into account.

Despite the need for this important further work, we hope that the results we reported here contributed to a better understanding of how properties shape conceptual knowledge both in human tasks and in computational models.

Addresses of the Authors
Marco Baroni, CIMeC (Centro Interdipartimentale Mente/Cervello), Università di Trento, c. Bettini 31, I-38068 Rovereto TN <marco.baroni@unitn.it>
Alessandro Lenci, Dipartimento di Linguistica, Università di Pisa, v. S. Maria 36, I-53126 Pisa <alessandro.lenci@ilc.cnr.it>

Notes

* We would like to thank Luis von Ahn for providing us with the ESP data, Ken McRae and colleagues for making their norms publicly available, Dominic Widdows and colleagues for the Infomap toolkit. We thank Eduard Barbu, Brian Murphy and Massimo Poesio for many interesting discussions and ideas, and for pointing out important resources and references, and Emiliano Guevara for useful feedback.

1 The norms can be downloaded from the *Psychonomic Society Archive of Norms, Stimuli, and Data* (http://www.psychonomic.org/archive).

2 http://www.espgame.org/

3 More precisely, the labels that are permanently associated with the images in the ESP collection are those that have been agreed on by n pairs of players, with n a "threshold of goodness" empirically fixed by the ESP designers.

4 http://www.natcorp.ox.ac.uk/

5 http://infomap-nlp.sourceforge.net/

6 The full patterns also inclue POS tags and lemmas (*from*/*IN*/*from_a*/*DT*/*a*), as well as morphological information about target and property (so that *layers from an onion* and *layer from an onion* produce different patterns because of the number difference in the property). These aspects are omitted for readability.

7 http://wacky.sslmit.unibo.it

8 All fruit names in the set could denote the corresponding trees, but at least from the NORMS responses it is clear that the single fruit sense is more salient (a cherry is red and sweet, etc.).

9 We would like to thank Brian Murphy for kindly providing us with these data.

Bibliographical references

VON AHN Luis & Laura DABBISH 2004. Labeling images with a computer game. In DYKSTRA-ERICKSON Elizabeth & Manfred TSCHELIGI (eds.). *Proceedings of ACM Conference on Human Factors in Computing Systems, CHI 2004*. 319-326.

BARSALOU Lawrence 2005. Situated conceptualization. In Henry COHEN and Claire LEFEBVRE (eds.). *Handbook of Categorization in Cognitive Science*. Amsterdam: Elsevier. 619-650.

BURGESS Curt & Kevin LUND 1997. Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes* 12. 1-34.

CREE George S., Chris MCNORGAN & Ken MCRAE 2006. Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology: Learning, Memory & Cognition* 32. 643-658.

FIRTH John R. 1957. *Papers in linguistics*. London: Oxford University Press.

FODOR Jerry 1998. *Concepts: Where cognitive science went wrong*. Oxford: Oxford University Press.

GLENBERG Arthur M. & Michael P. KASCHAK 2002. Grounding language in action. *Psychonomic Bulletin & Review* 9. 558-569.

HARRIS Zellig 1968. *Mathematical structures of Language*. New York: Wiley.

HEARST Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. *Proceedings of COLING 1992*. 539-545.

KINTSCH Walter 2001 Predication. *Cognitive Science* 25. 173-202.

LANDAUER Thomas K. & Susan T. DUMAIS 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104(2). 211-240.

LEVI Judith N. 1978. *The syntax and semantics of complex nominals*. Academic Press: New York.

MARTIN Alex 2007. The representation of object concepts in the brain. *Annual Review of Psychology* 58. 25-49.

MCRAE Ken, George CREE, Mark SEIDENBERG & Chris MCNORGAN 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods* 37. 547-559.

MEYER David, Achim ZEILEIS & Kurt HORNIK 2003. Visualizing independence using extended association plots. In HORNIK Kurt, Friedrich LEISCH & Achim ZEILEIS (eds.). *Proceedings of DSC 2003*, online at http://www.ci.tuwien.ac.at/Conferences/DSC-2003/.

MEYER David, Achim ZEILEIS & Kurt HORNIK 2006. The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software* 17(3).1-48.

MURPHY Gregory 2002. *The big book of concepts.* Cambridge: The MIT Press.

PADÓ Sebastian & Mirella LAPATA. 2007. Dependency-based Construction of Semantic Space Models. *Computational Linguistics* 161-199.

POESIO Massimo & Abdulrahman ALMUHAREB. 2008. Extracting concept descriptions from the Web: The importance of attributes and values. In BUITELAAR Paul & Philipp CIMIANO (eds.) *Bridging the Gap between Text and Knowledge*. Amsterdam:105 Press. 29-44.

RAPP Reinhard 2003. Word sense discovery based on sense descriptor dissimilarity. *Proceedings of the Ninth Machine Translation Summit*, online at http://www.amtaweb.org/summint/MTSummit/FinalPapers

RAPP Reinhard 2004. A freely available automatically generated thesaurus of related words. In *Proceedings of LREC 2004*. 395-398.

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

SALOMON Karen O. & Lawrence W. BARSALOU 2001. Representing properties locally. *Cognitive Psychology* 43. 129-169.

VIGLIOCCO Gabriella, David P. VINSON, William LEWIS & Merrill GARRETT 2004. Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology* 48. 422-488.

VIGLIOCCO Gabriella & David P. VINSON 2007. Semantic representation. In M. Gareth GASKELL (ed.). *The Oxford Handbook of Psycholinguistics*, Oxford: Oxford University Press. 195-215.

VINSON David P., Gabriella VIGLIOCCO, Stefano CAPPA & Simona SIRI 2003. The breakdown of semantic knowledge along semantic field boundaries: Insights from an empirically-driven statistical model of meaning representation. *Brain and Language* 86. 347-365.

WU Ling-Ling & Lawrence BARSALOU. In press. *Perceptual simulation in conceptual combination: Evidence from property generation.* To appear in Acta Psychologica.

ZEILEIS Achim, David MEYER & Kurt HORNIK 2005. *Residual-based shadings for visualizing (conditional) independence.* Technical report 20, Department of Statistics and Mathematics, Wirtschaftsuniversität, Vienna, online at http://epub.wu-wien.ac.at/dyn/openURL?id=oai:epub.wu-wien.ac.at:epub-wu-01_871.

A. Concepts and concept categories

Target items used for property analysis, together with their categories.

Word	*Semantic Category*
chicken	bird-animal-natural
duck	bird-animal-natural
eagle	bird-animal-natural
owl	bird-animal-natural
peacock	bird-animal-natural
penguin	bird-animal-natural
swan	bird-animal-natural
cat	groundAnimal-animal-natural
cow	groundAnimal-animal-natural
dog	groundAnimal-animal-natural
elephant	groundAnimal-animal-natural
lion	groundAnimal-animal-natural
pig	groundAnimal-animal-natural
snail	groundAnimal-animal-natural
turtle	groundAnimal-animal-natural
banana	fruit-vegetable-natural
cherry	fruit-vegetable-natural
pear	fruit-vegetable-natural
pineapple	fruit-vegetable-natural
corn	green-vegetable-natural
lettuce	green-vegetable-natural
mushroom	green-vegetable-natural
onion	green-vegetable-natural
potato	green-vegetable-natural
bottle	tool-artifact
bowl	tool-artifact
chisel	tool-artifact
cup	tool-artifact
hammer	tool-artifact
kettle	tool-artifact
knife	tool-artifact
pen	tool-artifact
pencil	tool-artifact
scissors	tool-artifact
screwdriver	tool-artifact
spoon	tool-artifact
telephone	tool-artifact
boat	vehicle-artifact
car	vehicle-artifact
helicopter	vehicle-artifact
motorcycle	vehicle-artifact
rocket	vehicle-artifact
ship	vehicle-artifact
truck	vehicle-artifact

B. Property classification scheme

Property classification scheme, adapted from Wu & Barsalou: submitted and McRae et al. (2005).

Class	*Property Type*	*Code*	*Example*
Taxonomy (c)	Coordinate	cc	*cat*-dog
	Superordinate	ch	*cat*-animal
Entity (e)	Associated abstract entity	eae	*telephone*-information
	Entity behavior	eb	*lion*-roar
	External component	ece	*truck*-wheel
	External surface property	ese	*banana*-yellow
	Internal component	eci	*car*-engine
	Internal surface property	esi	*pineapple*-crunchy
	Larger whole	ew	*cow*-cattle
	Made-of	em	*bottle*-glass
	Quantity	eq	*pear*-slice
	Systemic feature	esys	*elephant*-wild
Situation (s)	Associated entity	se	*spoon*-bowl
	Associated event	sev	*watermelon*-picnic
	Function	sf	*scissors*-cut
	Action	sa	*banana*-eat
	Location	sl	*ship*-port
	Participant	sp	*boat*-fisherman
	Time	st	*pineapple*-summer
Introspective (i)	Cognitive operation	io	*snail*-like a slug
	Evaluation	ie	*pineapple*-delicious
	Negation	in	*penguin*-cannot fly

C. Results

Raw property type counts for each target space.

Property Type	*NORMS*	*ESP*	*SVD*	*StruDEL*
cc	3	51	112	13
ch	43	14	16	32
eae	0	0	1	6
eb	37	8	1	39
ece	97	40	13	16
eci	12	5	8	9
em	24	6	1	1
eq	0	0	0	1
ese	57	30	7	9
esi	12	0	1	1
esys	20	0	9	5
ew	0	0	4	5
ie	3	0	1	0
in	3	0	0	0
io	1	0	0	0
Out	0	128	49	27
sa	5	8	26	93
se	8	47	147	43
sev	0	4	8	6
sf	70	37	20	61
sl	28	43	8	44
sp	15	4	7	10
st	1	1	1	1