

Multimodal word meaning induction from minimal exposure to natural text

Angeliki Lazaridou^a, Marco Marelli^b, Marco Baroni^a

^a*Center for Mind/Brain Sciences, University of Trento*

^b*Department of Experimental Psychology, Ghent University*

Abstract

By the time they reach early adulthood, English speakers are familiar with the meaning of thousands of words. In the last decades, computational simulations known as distributional semantic models have demonstrated that it is possible to induce word meaning representations solely from word co-occurrence statistics extracted from a large amount of text. However, while these models learn in batch mode from large corpora, human word learning proceeds incrementally after minimal exposure to new words. In this study, we run a set of experiments investigating whether minimal distributional evidence from very short passages suffices to trigger successful word learning in subjects, testing their linguistic and visual intuitions about the concepts associated to new words. After confirming that subjects are indeed very efficient distributional learners even from small amounts of evidence, we test a distributional semantic model on the same multimodal task, finding that it behaves in a remarkable human-like way. We conclude that distributional semantic models provide a convincing computational account of word learning even at the early stages in which a word is first encountered, and the way they build meaning representations can offer new insights into human language acquisition.

Keywords: word learning, distributional semantics, language and the visual world, one-shot learning, multimodality

1. Introduction

Humans know the meaning of a huge number of words. A high-school-educated English speaker possesses a vocabulary of as many as 60,000 terms (Aitchison, 1993). While it hits its peak early on, word learning continues throughout life (Bloom and Markson, 1998). Given that only a small portion of words are acquired through explicit instruction, we must be very good at inferring meaning from the contexts in which they occur. Since the pioneering work of Landauer and Dumais and others in the nineties (based on a much older tradition of psychological studies confirming the importance of context in word learning, see, e.g., Miller and Charles, 1991; Nagy et al., 1987; Sternberg

and Powell, 1983; Werner and Kaplan, 1950, among many others), a number of computational simulations have shown that, indeed, algorithms solely relying on patterns of word co-occurrence in large amounts of text (or *corpora*) induce distributed word meaning representations that are remarkably apt at capturing various aspects of human semantic competence. For example, they can pass a challenging synonym identification test (Landauer and Dumais, 1997), predict semantic priming (Padó and Lapata, 2007) and model the selectional preferences of specific verbs (Erk et al., 2010). These *distributional semantic models* (DSMs) have however until now been trained in *batch* mode, and evaluated after they have processed in full a large input corpus. Even simulations focusing on incremental learning tested performance changes in large steps (e.g., blocks of 100,000 words of running text in Baroni et al., 2007). Humans, however, learn words one-by-one in an incremental fashion, and they can learn the meaning of new words from very limited exposure. Often, a single encounter suffices (Trueswell et al., 2013). Can the distributional mechanisms shown by DSM simulations to be so effective in the long run also play a role in the initial stages of word learning from limited linguistic context?

There are *a priori* reasons to be doubtful that this would be the case. By its very nature, distributional learning proceeds by progressive accumulation of association statistics. Single co-occurrences are not expected to be very telling, and it is only after extended periods of linguistic exposure that robust distributional patterns are expected to emerge. This is rather problematic if one wants to claim that distributional learning is a plausible mechanism employed by humans to learn word meaning, as it suggests that most day-to-day word learning must happen by different mechanisms, and only at a much later stage (and only for sufficiently frequent words) whatever has already been learned about a word is complemented with evidence from long-run distributional learning. Distributional learning thus starts looking like it is largely redundant with respect to other, more fundamental mechanisms. Indeed, *one-shot learning*, our ability to learn from just one example of a new word or concept, is often mentioned as a strong objection against the cognitive plausibility of distributional and more generally associationist learning mechanisms (e.g., Lake et al., 2016, Trueswell et al., 2013). We think that this conclusion might be premature. Since discourses are generally coherent, even single word co-occurrences might often be at least broadly informative about a word meaning (a factor we will quantify below), making small-sample distributional learning generally reliable. Moreover, as a learner is progressively exposed to more language, we might observe a distributional bootstrapping effect, such that known words occurring as contexts of a new word help establishing a reasonable distributional profile for the new term quickly. Thus, in this study we decided to tackle, both experimentally and through a computational DSM simulation, the issue of whether it is possible for a learner to induce a reasonable semantic representation of a word from just very few (minimally 2, maximally 6) occurrences of the word in sentence contexts sampled to be representative of what one would encounter by chance in natural text (the minimum is set to 2 occurrences instead of 1 for technical reasons related to how we build our stimuli). Our results suggest that subjects

(and DSMs) can indeed come up with at least a basic representation of a word meaning from this minimal amount of purely distributional evidence. We take this to be an important result that, combined with the many studies that have shown how distributional models of meaning naturally account for many linguistic and psychological phenomena, both at the aggregate and at the item level, brings strong support for the view that distributional learning, specifically as implemented in DSMs, constitutes a plausible (although by no means unique) mechanism explaining our remarkable word learning capabilities.

Despite the theoretical importance of the issue, there is relatively little work focusing on fast learning of new word meanings from limited contexts. McDonald and Ramscar (2001) proved that the distributional characteristics of short made-up linguistic contexts surrounding rare or nonce words significantly affect subjects' semantic similarity intuitions about these words. Recent ERP work has moreover shown that subjects can learn novel word meanings from minimal exposure to artificial sentences containing the words. Borovsky and colleagues (2010; 2012) found that a single exposure to a novel word within an informative sentence context (i) suffices to trigger plausibility effects (signaled by N400 patterns) concerning the felicity of the new word as a grammatical object of an existing verb (for example, after being presented with "*He tried to put the pieces of the broken plate back together with marf*", subjects displayed a semantic violation effect for "*She drove the marf*"); and (ii) the novel word primes a semantically related target (as measured by reduction in N400 amplitude) just like a known word would. Using similar methods, Mestres-Missé et al. (2007) showed that, when a nonce word is presented in highly informative sentence contexts, three exposures suffice for the word to prime semantically related terms in a way that is indistinguishable from the effect of a comparable real word. In all these experiments, contextual effects were triggered by embedding the novel word in carefully constructed sentences. Thus, the results constitute proof of principle that, in the presence of highly informative surrounding words, subjects can learn new word meanings from minimal exposure, but they do not establish the result beyond artificial contexts. Put differently, they do not tell us whether real linguistic contexts, of the sort that one would encounter during everyday reading or conversation, are informative enough for subjects to be able to exploit them for similarly rapid word learning.

There are two further potential shortcomings that often affect word learning studies. First, learning a new word is not the same as learning a new *meaning*. It is easy to think about situations in which we acquire new words for familiar meanings (synonym learning). We may find out, for example, that the *buffalo* is also called *bison* (actually, the latter is a more proper term). One can also attach new meanings to familiar words, as in the case of polysemy. We may have originally known, for example, that *spam* denotes a type of canned meat, and only subsequently extended the word to denote junk e-mail. However, arguably, most often a speaker must learn how *new* words refer to *new* meanings (e.g., learning the word *iPad* the first time you heard about it, or saw one). That is, new word and concept acquisition typically proceed in parallel. Still, in much of the previous literature on word learning, including the studies we reviewed,

participants are *de facto* asked to associate new labels to familiar meanings, that is, they are faced with a synonym learning task. In the example above, subjects must discover that *marf* is a new synonym of *glue* (Mestres-Missé et al., 2007, report that, in 91% of their informative-condition sentences, subjects were able to guess the real word masked by the nonsense string, or a closely related term).

Second, the relevant literature follows two main strategies to verify successful learning. A classic approach is to test if a subject can associate a new word with the right visual referent: e.g., the subject hears “Oh look, a *marf*” while seeing images of glue and a number of distractors, and the experiment measures whether she directs her gaze at glue (Trueswell et al., 2013). Being able to pick the right referent is a fundamental aspect of learning the meaning of a (concrete) word, but, as Borovsky et al. (2012) observe, “our lexical knowledge is often far richer than simple associations between labels, physical objects and features. Word representations are complex and multi-faceted. [...] The] meaning [of a novel word . . .] must be appropriately situated within the local context and dynamic semantic landscape of the mental lexicon.” The studies discussed above address this criticism by probing purely language-based semantic properties of the novel word (e.g., whether it primes a related term). However, this method, in turn, gives up the link to visual referents as an important component of meaning. Especially when investigating the role of purely linguistic distributional evidence, it is fundamental instead to test whether subjects are not only getting lexical relations right (which might be explained away by relatively superficial textual co-occurrence effects), but also inferring the likely perceptual properties of the concept they are learning about.

In the present study, we adopt a novel experimental paradigm that is aimed at addressing the concerns above. First, we do not employ artificially constructed texts to probe word learning. We extract natural sentences from existing corpora, substituting a target word with a non-word. Second, inspired by the perception literature, in which “chimeric” images are generated by mixing visual features of different objects or animals (Sarri et al., 2011), we try to trigger genuine novel-meaning learning (as opposed to synonym matching) by mixing sentences referring to two related but distinct words. For example, the context of one of our lexical chimeras is a small random sample (that we call a *passage*) of sentences originally containing either the word *bear* or the word *gorilla*, both replaced by the non-word *mohalk*. Because gorillas and bears are related, the passage will paint a coherent picture of mohalks as large mammals, but subjects won’t be able to play a synonym guessing game, since there is no single word that *mohalk* is masking. Third, participants are asked to evaluate a series of probe items for their similarity to the novel term meaning, thus going beyond the simple labeling tasks criticized by Borovsky and colleagues. However, since we aim at testing both strictly linguistic and more general conceptual knowledge that subjects associate to the novel word, we employ as probe items both words (Experiment 1), and images (Experiment 2) of objects that are expected to be more or less related to the chimeric concept. Finally, after measuring human performance on our distributional learning task, we simulate it with a novel DSM architecture combining linguistic and visual knowledge.

Our results confirm that (i) (adult) subjects can extract basic knowledge about the meaning of a new word, including both linguistic and visual aspects of the denoted concept, from a very small number of sentences randomly extracted from natural texts, and that (ii) a DSM exposed to the same kind of limited evidence induces multimodal knowledge about the new word in a way that is remarkably similar to what we observe in subjects.

The rest of the article is organized as follows: Section 2 describes our experimental materials and introduces the computational model. The two experiments are reported in sections 3 and 4, respectively. The final discussion in Section 5 summarizes our main results, and looks at our computational learning method from the broader perspective of word learning models.

2. Experimental setup

2.1. Experimental materials

We investigated the generation of novel concepts from short passages in adult human participants (see, e.g., Gillette et al., 1999, for general motivation to study word learning in adults). In particular, we created novel concepts, that we call *chimeras*, by combining two related but distinct words, the *chimera components*. A *passage* then consisted of a mixture of natural corpus sentences (minimally 2, maximally 6) that contained the two components, except that all instances of either component were masked by the same nonce word. Subjects were induced to believe that the sentences all pertained to the same novel concept, and they were tested for their intuitions about the meaning of this novel concept by being asked relatedness judgments between the latter and other terms, that we call *probes*. For example, Figure 2 below shows materials pertaining to the chimera with *gorilla* and *bear* as components. In one trial, illustrated in the figure, the shared nonce word masking the components was *mohalk*. Subjects (and the computational model) were presented a passage composed of 4 sentences where instances of *gorilla* and *bear* were replaced by *mohalk*, and immediately asked to rate the degree of relatedness of *mohalk* with, e.g., *dishwasher* or (in a separate trial) *lion*. As the figure shows, the probes were presented verbally in Experiment 1 and visually in Experiment 2.

2.1.1. Chimeras

To obtain the chimeras, we started with 33 basic-level concrete concepts that were also used by Frassinelli and Keller (2012) (three of their concepts were excluded due to technical reasons). We call each of these concepts a *pivot*, and we match it with another concept, that we call a *compatible* term, by using the following procedure. For each pivot, we ordered all the terms in the norms of McRae et al. (2005) by similarity to the pivot, using the pre-compiled similarity table available with these concept norms (which is based on conceptual feature overlap, where concept features were elicited from subjects). We traversed the resulting ranked lists, picking as compatible term the first word that was not a synonym, close co-hyponym or hyper-/hyponym of the

pivot. Whenever possible, we tried to match pivot and compatible terms with similar usage (e.g., both mass terms, or both likely to frequently occur in the plural) and reasonably similar in visual shape and function. No pivot was picked as compatible term of another pivot, and no compatible term was repeated across pivots. Except for *plane/ship*, the compatible term was always among the 10 most similar items to the pivot (for *plane*, the most similar concepts were birds). A chimera is given by the combination of a pivot and the corresponding compatible term (the chimera components). The first two columns of Table 1 report the full list of matched components forming our chimeras.

2.1.2. Probes

Each chimera was associated with 6 words, whose degree of relatedness to the chimeras had to be assessed by the subjects based on the limited contexts surrounding the latter. We refer to such words as probe terms. We obtained probes spread across the relatedness spectrum by sampling them from different bins based on averaged McRae-norms similarity to chimera components. Examples of probes sampled in this way include *turnip* for the *corn/yam* chimera (in the top similarity bin according to the McRae norms) and *chipmunk* for *toaster/microwave* (bottom similarity bin).

In order to measure the degree of success of subjects and computational model in positioning a chimera in the right zone of conceptual space, we needed ground-truth chimera-probe relatedness (CPR) judgments. Since the chimeras are novel concepts, we estimated this quantity indirectly, by averaging similarity ratings between the (explicitly presented) chimera components and the probes, produced by a control group in a preliminary crowdsourcing experiment (Schnoebelen and Kuperman, 2010), using the Crowdfunder platform.¹ The ratings for the two components of a chimera with each probe were collected separately. Subjects in the control group were given no cue that we intended to combine such ratings. Subjects were asked to rate the semantic relatedness of each chimera component word to the corresponding probes, for a total of 396 pairs (66 components by 6 probes), using a 7-point scale ranging from 1 (“completely unrelated”) to 7 (“almost the same meaning”). Ten judgments were collected for each pair. In the instructions, we stressed that we were interested in intuitions concerning word *meanings*, and that the relation between the element pairs could be more or less strict. The CPR of a chimera-probe pair was computed as the average rating of the two chimera components with the probe. For example, given the *bear/gorilla* chimera and the *lion* probe, we used averaged *bear-lion* and *gorilla-lion* ratings obtained from the control group as our CPR estimate.

Since participants’ semantic intuitions about chimeras in the experiments below were elicited through short text passages that pertain, in equal amounts, to the two components, it is reasonable to assume that the full chimera concept is an average of the component meanings (the *bear/gorilla* chimera denotes an

¹<http://www.crowdfunder.com>

animal somewhere in-between bears and gorillas), and that the “real” CPR score should be close to that obtained by averaging component-probe ratings. Note that one chimera’s components are very similar concepts. Consequently, they will in general have very comparable similarities to the probes (correlations between ‘component’-specific CPRs across all chimeras and probes are quite high: $r = .73$ for word-to-word judgments and $r = .69$ for the word-to-image judgments we use in Experiment 2 as discussed next). It is thus unlikely that the CPR scores of a chimera as a genuine separate concept would be very different from the means of the two component CPR scores. If subjects find both *bears* and *gorillas* very different from a *dishwasher*, it is probable that they would assign a similarly low *dishwasher* score to the *bear/gorilla* chimera, if such creature existed. Relatedness scores *averaged* across the two chimera components worked better as CPR estimates (in the sense of fitting our experimental results more closely) than relying on maximum or minimum component-to-probe relatedness. Even using the ground-truth scores for the most informative component in each passage (in the sense discussed below) did not lead to an improvement in model fit. To conclude, although averaging component CPR scores can only provide an indirect estimate of a chimera’s CPR score, we have good reasons to believe that such estimate is quite accurate.

For Experiment 2, we replaced the probe words with images of the corresponding concepts. As seen in Figure 2 below, images were not artificially edited to represent the probe concepts in stereotypical settings. They were instead natural pictures of the relevant objects (that is, they were shot by amateur or professional photographers for their own purposes), taken from the ImageNet database we will briefly describe in Section 2.2.2.

Table 1 presents, for each chimera, its components together with the 6 probes and the respective control-group-generated word- and image-based CPR values. For example, the first row reports that one chimera was formed by the *alligator* pivot and by the compatible term *rattlesnake*, that the McRae similarity between these concepts is 0.39 (on a 0-1 scale), that *rattlesnake* is the second most similar term to *alligator* in the McRae norms (thus, it has rank 2 in the sorted neighbour list). It further shows that the *crocodile* probe received an average CPR of 4.15 (on a 1-7 scale) with this chimera when *crocodile* was presented as a word, and of 4.20 when it was presented as a picture, and similarly for all the other probes associated to the chimera.

—————Insert Table 1 about here —————

2.1.3. Passages

Each chimera was associated to 10 passages consisting of 6 sentences each (3 sentences per component). The latter were randomly extracted, in comparable proportions, from two corpora representative of written and spoken English (British National Corpus)² and Web texts (ukWaC, see Section 2.2.2 below), respectively (we avoided other commonly used textual sources likely to provide

²<http://www.natcorp.ox.ac.uk>

Pivot	Compat.	Sim(Rank)	Probes					
			Word-based CPR			Image-based CPR		
alligator	rattlesnake	0.39 (2)	crocodile	iguana	gorilla	banner	buzzard	shovel
bomb	missile	0.64 (1)	4.15 4.20	2.95 2.65	1.75 1.55	1.15 1.00	2.10 1.40	1.05 1.05
broccoli	spinach	0.74 (1)	bazooka	gun	bayonet	bullet	buckle	canoe
cannon	rifle	0.43 (2)	3.60 3.25	3.10 2.55	2.40 2.50	3.85 3.10	1.20 1.10	1.20 1.00
car	van	0.60 (1)	celery	radish	grape	salamander	budgie	pot
caterpillar	cockroach	0.33 (1)	2.80 2.70	2.85 2.30	2.45 1.80	1.40 1.00	1.45 1.00	1.15 1.70
cello	bagpipe	0.45 (8)	pistol	bomb	harpoon	trolley	lion	bagpipe
clarinet	trombone	0.67 (2)	3.45 2.80	3.35 2.55	2.55 3.55	1.00 1.20	1.20 1.15	1.20 1.20
corkscrew	grater	0.38 (5)	skateboard	jeep	train	fridge	shed	parakeet
corn	yam	0.29 (6)	1.70 1.20	5.10 6.05	3.20 2.35	1.25 1.05	1.30 1.05	1.30 1.00
cucumber	celery	0.66 (1)	beetle	grasshopper	spider	shack	porcupine	crane
dishwasher	oven	0.37 (4)	2.90 2.45	3.30 2.85	3.30 2.20	1.20 1.20	1.80 1.80	1.55 1.45
drums	tuba	0.55 (2)	harmonica	drum	racquet	cabin	house	bolt
elephant	bison	0.46 (5)	2.75 1.80	2.55 1.75	1.20 1.10	1.25 1.15	1.15 1.00	1.30 1.00
gorilla	bear	0.42 (5)	banjo	gorilla	whistle	worm	dress	pine
guitar	harpsichord	0.76 (3)	2.85 2.60	1.15 1.00	2.85 1.60	1.25 1.20	1.20 1.10	1.25 1.00
harp	banjo	0.75 (3)	1.55 1.55	1.90 1.25	1.10 1.15	1.10 1.25	1.35 1.15	1.40 1.10
kettle	pan	0.23 (8)	turnip	eggplant	parsley	peach	buffalo	pickles
ladle	colander	0.51 (3)	3.20 2.50	3.10 2.25	2.80 1.60	2.60 1.80	1.20 1.35	1.15 1.05
mittens	socks	0.51 (3)	rhubarb	onion	pear	strawberry	limousine	buggy
owl	partridge	0.61 (2)	3.15 2.40	3.05 2.45	2.45 1.85	2.60 2.45	1.20 1.00	1.25 1.05
peacock	goose	0.70 (2)	stove	microwave	kettle	cage	wastebin	cushion
piano	accordion	0.61 (3)	3.60 4.70	3.85 3.40	2.10 1.75	1.15 1.00	1.45 1.10	1.25 1.05
plane	ship	0.20 (36)	bagpipe	harmonica	whistle	shotgun	bear	leopard
potato	turnip	0.54 (1)	3.05 2.60	2.70 1.90	2.65 1.60	1.20 1.00	1.05 1.10	1.20 1.00
refrigerator	closet	0.38 (4)	caribou	groundhog	hare	spider	catapult	bouquet
saxophone	harmonica	0.49 (8)	3.00 2.40	1.90 1.80	2.15 1.45	2.05 1.25	1.10 1.10	1.10 1.00
scarf	sweater	0.47 (2)	tuba	racquet	barrel	shack	raisin	bolts
scooter	skateboard	0.50 (5)	2.65 2.35	1.30 1.10	1.25 1.05	1.25 1.00	1.25 1.00	1.15 1.05
toaster	microwave	0.55 (3)	1.30 1.10	1.25 1.05	1.25 1.05	1.40 1.05	1.35 1.05	1.05 1.00
train	bus	0.35 (3)	3.80 2.95	3.20 1.95	1.25 1.00	1.40 1.05	1.35 1.05	1.10 1.00
trouser	shirt	0.35 (2)	2.25 1.70	2.45 1.30	1.25 1.10	1.15 1.05	1.05 1.05	1.20 1.10
violin	flute	0.50 (6)	tongs	corkscrew	kettle	stool	pistol	helmet
			2.70 1.85	1.55 1.40	2.50 1.40	1.10 1.05	1.25 1.00	1.10 1.10
			2.15 1.50	2.35 1.80	2.25 1.85	1.25 1.80	2.20 1.35	1.20 1.00
			2.60 2.85	3.00 2.00	1.20 1.10	1.85 1.40	2.05 1.80	1.20 1.05
			3.30 1.90	1.40 1.10	2.10 1.60	2.35 1.45	2.25 1.00	1.45 1.00
			harpsichord	trumpet	typewriter	penguin	olive	lettuce
			3.70 4.10	3.20 2.10	1.50 1.45	1.20 1.10	1.05 1.05	1.20 1.10
			jet	yacht	hawk	stork	corkscrew	nightgown
			4.90 4.45	3.65 3.15	1.25 1.10	1.30 1.30	1.25 1.05	1.10 1.15
			broccoli	onion	tomato	trout	cantaloupe	cork
			2.90 1.95	3.05 2.30	2.85 2.25	1.70 1.30	2.30 3.00	1.20 1.00
			cupboard	basement	mixer	dishwasher	ladle	boat
			3.15 2.60	1.65 1.40	1.85 1.20	1.85 2.35	1.40 1.00	1.30 1.05
			clarinet	harp	buckle	wrench	urn	mackerel
			3.60 2.75	3.25 2.25	1.20 1.00	1.05 1.00	1.45 1.10	1.15 1.05
			glove	robe	tie	swimsuit	nylons	spear
			3.25 2.30	2.95 2.40	3.35 2.15	2.05 1.05	2.55 1.40	1.30 1.00
			cart	jeep	boat	toy	pencil	tomato
			2.35 1.30	2.15 1.15	1.65 1.40	3.65 1.00	1.00 1.00	1.05 1.00
			pot	apron	kettle	tongs	cheetah	chipmunk
			1.90 1.85	1.50 1.50	2.35 2.05	1.85 1.30	1.15 1.05	1.20 1.05
			jet	taxi	buggy	submarine	crane	grasshopper
			2.70 2.05	2.70 2.45	2.35 2.45	2.15 1.35	1.40 1.60	1.10 1.00
			pants	shawl	cape	curtain	pajama	cart
			4.25 4.15	2.90 1.60	2.50 2.55	1.50 1.20	3.35 3.90	1.15 1.15
			harp	drum	racquet	colander	rocket	radish
			3.20 2.55	2.75 2.05	1.25 1.00	1.15 1.05	1.20 1.15	1.20 1.05

Table 1: Chimeras and probes. For each chimera, the table reports, in this order, the chimera pivot and compatible components, their McRae similarity (and rank of compatible item among neighbours of pivot) and the probes with (left) word-based and (right) image-based CPR scores.

overly informative text, such as the Wikipedia). The sentences were manually checked to make sure they did not predicate mutually contradictory properties of the chimera (as the chimera components are related but different concepts). We also substituted sentences in which a component word was not used as intended (e.g., *bear* used as a verb). The sentences were not edited in any way. On average, they contained 17.6 words each. All occurrences of either original component in a passage were replaced with the same non-word (for example, all the occurrences of *gorilla* and *bear* in a given passage were replaced by *mohalk*). The 330 non-words (33 chimeras by 10 passages) were generated using WUGGY (Keuleers and Brysbaert, 2010). They were 2- or 3-syllable-long nonsense strings respecting English orthotactics, and not containing productive affixes (e.g., *-ing*).

In order to quantify the amount of context that is needed to learn new meanings, we manipulated the number of sentences by creating shorter passages from the original ones. Three possible *passage length* levels were considered (2, 4, 6 sentences), each of them including an even number of sentences for either component (incidentally, this is the reason why we can't go down to 1 occurrence per chimera).

We calculated an *a-posteriori* estimate of passage informativeness with respect to the chimeras they contain. A passage is informative if it contains words that are descriptive of the chimera component concepts: e.g., a *bear/gorilla* passage containing words such as *fur* or *animal*, that are likely to be used in definitions or prototypical descriptions of bears and gorillas, is more informative than a passage that does not contain such words. Concretely, the informativeness score of a passage is the number of distinct words in the passage that are identified as properties of the component concepts in the conceptual norms of McRae et al. (2005). We checked that informativeness was distributed not only across different passages, but also within each chimera. The average informativeness range was 0-4.5, going from *car-van* (with an informativeness range of 0-2) to *caterpillar-cockroach* and *guitar-harpsicord* (with an informativeness range of 0-7). Moreover, the correlation between passage length and informativeness was relatively low ($r = .38$), indicating that one variable could not be reduced to the other.

One possible concern with our design is that many passages could be so informative about the chimera components that subjects could have guessed these items and based their ratings on them, rather than trying to infer a genuinely new meaning from the passage. To address this concern, we ran an additional crowdsourcing experiment via Crowdfunder. In this control experiment, we asked subjects if they could guess the word or words that were masked in each of the 330 6-sentence passages. A total of 101 subjects took part in the experiment. Each passage was presented to 10 subjects, and a single subject could maximally rate 50 passages. Note that, whereas the subjects of our main experiments, to be introduced below, were instructed to think of the passages as involving a new concept, the current control group was explicitly instructed to try to guess existing words for the masked slots. Moreover, unlike the main experiment subjects, the control group was told that different sentences in the

same passages could involve different masked words, and they were allowed to list more than one word per passage. Given these differences, this control experiment is testing a worst-case scenario, in which no main-experiment subject followed our instructions, they decided to use a word-guessing strategy instead, and they also figured out that the passages might refer to different concepts. Despite the facilitated setup, for only 17% of the passages more than half of subjects were able to guess at least one of the two chimera components. However, in many of these cases the subjects guessed a chimera component simply because they produced a list of related concepts, that included the component. For example, for an *harmonica/saxophone* passage, a subject wrote *harp, guitar, bass, saxophone*, suggesting he/she understood that the passage referred to a musical instrument, but he/she was not really able to tell which one. But this is exactly the sort of effect we are after, with subjects being able to tell that the “new” *harmonica/saxophone* chimera is an instrument, without being able to map it precisely to an existing one. The proportion of passages in which more than half the subjects were able to guess at least one chimera component without also producing irrelevant concepts is 10%. There is no single passage for which more than half the subjects were able to guess both chimeras.

2.2. Computational model

By relying on the observation that words occurring in similar contexts tend to have similar meanings, DSMs exploit the co-occurrence patterns of words with corpus contexts to derive distributed (vector-based) meaning representations. Geometric similarity measures (e.g., the cosine of the angle formed by vectors representing different words) are then used to approximate the degree of semantic relatedness between words.

We adopt here the *multimodal skip-gram* model (MSG) that we have recently developed (Lazaridou et al., 2015a). Related approaches have been explored in the literature, e.g., by Howell et al. (2005) and Kievit-Kylar and Jones (2011), but to the best of our knowledge MSG is the first model of its kind that exploits genuine visual data, instead of manually-crafted symbolic proxies for them. MSG is a multimodal extension of the skip-gram model proposed by Mikolov et al. (2013).

Figure 1 illustrates a MSG learning step. MSG processes input sentences one-by-one, similarly to how humans are exposed to language (by attending a conversation, reading a book, etc). In this way, the meaning of words is gradually induced by constantly updating both their target and contextual representations based on the evidence that each new sentence provides. MSG simulates *grounded* language learning by combining linguistic and visual contexts. In practice, for a limited number of concrete words, we expose our model not only to the sentences in which they occur in the source corpus, but also to a visual representation of the object denoted by the word, obtained through automated visual feature extraction techniques from natural images containing the object. In the example in the figure, the *cat* word vector is updated to jointly predict linguistic context and visual properties that characterize cats. Intuitively, this

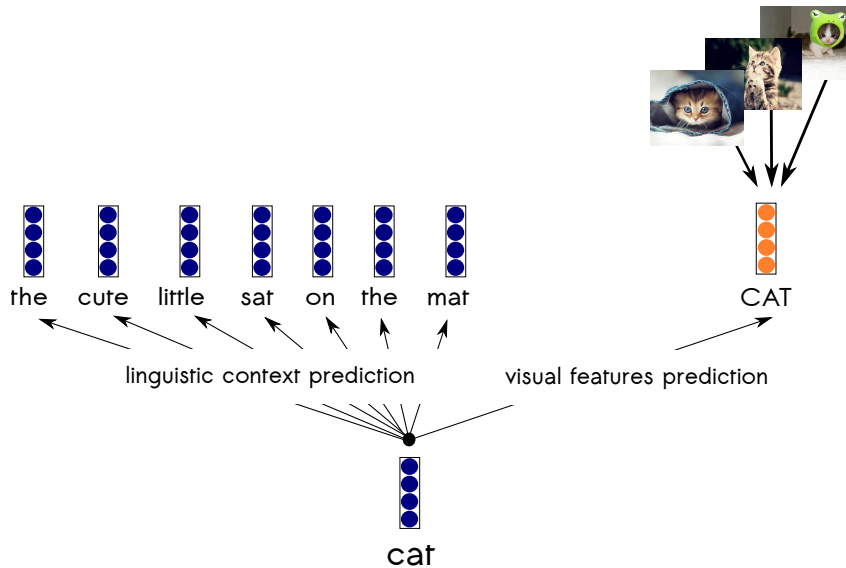


Figure 1: Illustration of a MSG learning step. When presented the sentence *the cute little cat sat on the mat*, the model updates the distributed representation of the word *cat* on the joint tasks of predicting the other words occurring in the sentence and visual features extracted from cat pictures. Distributed representations of the context words to be predicted are induced in parallel with the representations of the target words, whereas visual features are fixed in advance. Visual features are only provided for a subset of words (and presented at each occurrence of such words in the corpus). For words that are not associated to visual features, only the linguistic context prediction objective is used. However, because of its joint-objective architecture, the model propagates visual information to all words through linguistic similarity (e.g., even if *kitty* is not associated to a visual representation, its vector will presumably become similar to that of *cat* on linguistic grounds, and consequently it will tend to predict similar visual features).

emulates the process of hearing or reading words accompanied by concurrent perceptual stimuli, e.g., hearing about *cats* while looking at them.

—Insert Figure 1 about here —

In Lazaridou et al. (2015a), we have shown that visual information also naturally “propagates” to the majority of words for which no direct perceptual evidence is present, thanks to their linguistic similarity to words whose vectors are exposed to direct visual information. The same study also showed that MSG is the state of the art on many standard DSM semantic and multimodal benchmarks. Lazaridou et al. (2016) further employed MSG in a pilot study modeling word learning from multimodal child-directed data. Thus, we are not working with an *ad-hoc* model designed specifically to address novel word learning from minimal contexts, but with a competitive DSM that, when tested on known words, provides a satisfactory account of adult semantic competence, and which seems to also be able to account for certain aspects of word learning in children.

2.2.1. Model definition

As the MSG model traverses the input corpus one sentence at a time, word vector representations are induced by optimizing a joint objective involving prediction tasks in the linguistic and visual modalities. In the former, for each word in the current sentence the model must perform linguistic context prediction. The vector of the target word is updated to better predict the c words surrounding it in the sentence. In the example of Figure 1, the *cat* vector will be updated to improve its prediction of the words $\{the, cute, little, sat, on, the, mat\}$. Note that words are also assigned vectors in their role of contexts, and that improving prediction consists in making the target word (e.g., *cat*) vector more similar to those of the context words. In this way, words occurring in similar contexts end up developing similar vectors, according to the tenets of distributional semantics. The visual objective is activated whenever a word in the sentence being processed is associated to a visual representation, and consequently the word vector is also updated to improve prediction of the relevant visual features (right side of Figure 1).

Concretely, the objective function of MSG includes two terms:

$$\frac{1}{T} \sum_{t=1}^T (\mathcal{L}_{ling}(w_t) + \mathcal{L}_{visual}(w_t)) \quad (1)$$

where T is an index ranging over positions in the input text corpus. The linguistic objective $\mathcal{L}_{ling}(w_t)$ is the same as in the skip-gram model of Mikolov et al. (2013). It aims at inducing word representations that are good at predicting the *context* words surrounding a *target* word. Mathematically, it maximizes:

$$\sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j}|w_t) \quad (2)$$

where subscripts index corpus positions and c is the size of the window around target w_t , determining the set of context words to be predicted by the induced representation of w_t . Following Mikolov et al., we also introduce a subsampling step randomly discarding context words as an inverse function of their frequency, controlled by hyperparameter s . This gives more weight to the prediction of rarer and typically more informative words. Probability $p(w_{t+j}|w_t)$ is computed as:

$$p(w_{t+j}|w_t) = \frac{e^{u'_{w_{t+j}} \cdot u_{w_t}}}{\sum_{w'=1}^W e^{u'_{w'} \cdot u_{w_t}}} \quad (3)$$

where u_w and u'_w are the context and target vector representations of word w respectively, and W is the size of the vocabulary. Due to the normalization term, Equation (3) requires $O(|W|)$ time complexity. A considerable speedup to $O(\log |W|)$, is achieved by using the frequency-based hierarchical version of Morin and Bengio (2005), which we adopt here.

If a corpus word is not accompanied by a visual representation, the visual term $\mathcal{L}_{visual}(w_t)$ of Equation (1) is set to 0. When a word is associated to

visual features (obtained as described below), the visual term tries to maximize the similarity between the visual feature vector and the target word vector *projected* onto visual feature space. This requires estimating a cross-modal mapping matrix \mathbf{M} from linguistic onto visual representations, jointly induced with linguistic word representations, which, given a word vector u_{w_t} , transforms it to its visual equivalent $z_{w_t} = \mathbf{M}u_{w_t}$. We maximize cosine similarity between visual and mapped word vectors through a max-margin framework commonly used in models connecting language and vision (Weston et al., 2010; Frome et al., 2013). More precisely, we formulate the visual objective $\mathcal{L}_{visual}(w_t)$ as:

$$\sum_{w' \sim P_n(w)} \max\{0, \gamma - \cos(z_{w_t}, v_{w_t}) + \cos(z_{w_t}, v_{w'})\} \quad (4)$$

where γ is the margin, z_{w_t} is the visual estimate of the target word representation we aim to learn, v_{w_t} is the corresponding visual vector (fixed in advance) and $v_{w'}$ ranges over visual representations of words (associated to visual features) randomly sampled from distribution $P_n(w_t)$ (set to uniform). These random visual representations act as negative samples, encouraging z_{w_t} to be more similar to its own visual representation than to that of other words. The number of negative samples is controlled by hyperparameter k . To avoid overfitting, we also add an L2 regularization term for \mathbf{M} to the overall objective (Equation 1), with its relative importance controlled by hyperparameter λ .

2.2.2. Model training

Simulating the acquisition of adult-like competence, we pre-train MSG on ukWaC, a text corpus of about 2 billion running words from random Web pages.³ We associate all corpus occurrences of 3,825 distinct concrete words (about 5% of the total word tokens) with their visual representations. These words are selected as follows: They must have an entry in ImageNet (see next paragraph), occur at least 500 times in ukWaC and have concreteness score ≥ 0.5 according to Turney et al. (2011). There were 5,100 words matching these constraints, but 25% of them (1,275) were set apart to estimate the mapping function described in the next section. In total, 116 out of the 155 probes and 50 out of the 66 chimera components were associated to visual representations during training.

Visual representations are obtained by randomly sampling 100 images labeled with the word of interest from ImageNet (Deng et al., 2009), a large database of “natural” pictures downloaded from the Web and annotated with words. We automatically extract a 4096-dimensional vector from each image using the Caffe toolkit (Jia et al., 2014), together with the state-of-the-art convolutional neural network of Krizhevsky et al. (2012). Convolutional neural networks, taking inspiration from biological vision, extract multiple layers of increasingly abstract features from images. Our vectors correspond to activation on the top (fc7) layer of the network, that captures complex, gestalt-like shapes

³<http://wacky.sslmit.unibo.it>

(Zeiler and Fergus, 2014). For computational reasons, we reduce the vector to 300 dimensions through Singular Value Decomposition. The visual representation of a word is simply the centroid of the vectors of the 100 images labeled with the word.

The following hyperparameters are fixed without tuning: word vector dimensionality: 300; subsampling: $s=0.001$; window size: $c=5$. The following hyperparameters are tuned on the text9 corpus:⁴ $k=5$, $\gamma=0.5$, $\lambda=0.0001$. The remaining model parameters are estimated by back-propagation of error via stochastic gradient descent on the input corpus and associated visual representations.

2.2.3. Simulating relatedness judgments

Given the passage in which a chimera occurs, simulating the chimera-to-probe (CPR) relatedness judgments requires the generation of a vector representation for the chimera (that, recall, from the participants' point of view, is a novel word only occurring in the passage). Taking inspiration from early work on contextualizing word meaning in DSMs (Schütze, 1997), we represent a chimera in a sentence context as the centroid of the MSG vectors of all other (known) words in the sentence. The centroids obtained from each sentence in a passage are normalized and averaged to derive the passage-based chimera vector. This approach can leverage more information (harnessing the full vectors of all words in the passage) than simply updating a new chimera vector in a single step of the MSG prediction task.

In Experiment 1, we quantify the degree of relatedness of chimera and probe word by measuring the cosine formed by the angle of the chimera vector (constructed as just described) and the pre-trained MSG vector of the probe word.

In Experiment 2, we need to measure the relatedness of the chimera with an *image* probe. To represent the probe, we extract a visual feature vector from the image presented to subjects with the same convolutional neural network used for MSG training (see previous section). Since the probe and the chimera vectors lay in different spaces (visual and linguistic, respectively), their mutual relatedness cannot be directly assessed as in Experiment 1. The two representations could be connected in two ways: by estimating a linguistic equivalent of the probe visual vector by means of a vision-to-language mapping function, or, alternatively, inducing a visual vector from the chimera linguistic vector through the inverse mapping. We report results obtained with the former procedure because the resulting relatedness scores were more in line with human judgments.

Specifically, we learn to map visual representations onto linguistic space. The mapping, consisting of weight matrix \mathbf{M}_{v2w} , is induced with a max-margin ranking loss objective, analogously to what is done in MSG estimation. The loss for pairs of visual and linguistic training items $(\mathbf{x}_i, \mathbf{y}_i)$ and the corresponding

⁴<http://mattmahoney.net/dc/textdata.html>

It was illustrated by a picture of Bernie Grant, the black Labour candidate,
with the hairy body of a **MOHALK**. [*gorilla*]
ANCHORAGE-- During September an unusually high number of **MOHALKS** were sighted
in Alaska's North Slope oil fields. [*bear*]
But watch out for nerds running up the down escalators, down the up escalators and generally
acting as **MOHALKS** on crack. [*gorilla*]
Of interest, during these cleaning activities he unearthed what appeared to be a **MOHALK**
trap wedged in a crack. [*bear*]

Experiment 1

DISHWASHER

LION

Experiment 2



Figure 2: Example of the materials used in an experimental trial for the *gorilla/bear* chimera, corresponding to the nonce string MOHALK (the original word in each sentence, shown in squared brackets, was not presented to subjects). Unrelated and related probes are shown bottom left and right, respectively (in Experiment 1, only word probes were presented, in Experiment 2, only images).

mapping-derived predictions $\hat{\mathbf{y}}_i = \mathbf{M}_{v2w}\mathbf{x}_i$ is defined as

$$\sum_{j \neq i}^k \max\{0, \gamma - \cos(\hat{\mathbf{y}}_i, \mathbf{y}_i) + \cos(\hat{\mathbf{y}}_i, \mathbf{y}_j)\} \quad (5)$$

where γ and k are tunable hyperparameters denoting the margin and the number of negative examples, respectively. The optimal hyperparameters $\gamma = 0.7$ and $k = 2$ were chosen on a set of 300 validation concepts from the training set. To train the mappings, we use word and visual representations derived from 1,275 concrete concepts (selected as described in the previous section). We estimate the mapping parameters \mathbf{M}_{v2w} with stochastic gradient descent and per-parameter learning rates tuned with AdaGrad (Duchi et al., 2011).

3. Experiment 1: Estimating relatedness of chimeric concepts to word probes

3.1. Methods

3.1.1. Participants

Participants were recruited through crowdsourcing, and in particular using the Crowdfunder platform. As is standard for Crowdfunder studies, participants were free to abandon a job at any time, and there were no restrictions on how many jobs a participant could perform. The subjects eventually participating in

the experiment were 213. As an average, each chimera was rated by 143 distinct subjects.

3.1.2. Procedure

Relatedness ratings between the chimeras in each passage and the associated probes were collected. Passage length and probes were counterbalanced across items using a Latin-square design, so that a given combination of passage and probe appeared only once in each list. Each of the resulting 18 lists (6 probes by 3 passage lengths) included 330 items (33 chimeras by 10 passages). Each list was administered as a separate Crowdfunder job, resulting in a full within-subject design (that is, participants were presented all possible levels of factorial predictors, and a representative distribution of continuous predictors). In order to avoid large familiarity effects between different lists, jobs were presented in separate days. Moreover, non-words were randomized across lists, so that each non-word was associated to a different passage in each list.

Participants were asked to judge how related a familiar word (the probe) was to the concept denoted by the unknown word in the presented sentence. An example trial is presented in Figure 2 (ignore the images). Participants were told that the unknown word could be thought of as referring to a new or very rare concept, or to a thing that might exist in the parallel universe of a science fiction saga. We also specified that the relation could be stricter or looser, and that we were specifically interested in the word *meanings*, rather than forms or sounds. Relatedness was rated on a 5-point scale. Seven judgments were collected for each item.⁵

—————Insert Figure 2 about here —————

3.1.3. Data analysis

Human ratings were analyzed through mixed-effects models (Baayen et al., 2008) including random intercepts for subjects and chimeras. The main predictor was CPR. In the analysis we also included covariates as potential modulators of this effect. We considered the interaction of CPR with passage informativeness. Moreover, passage length was also made to interact with CPR, to assess whether the raw amount of contextual information contributes to the quality of the generated novel representation. All predictors were mean-centered. We started from a full factorial model, and removed effects that did not contribute significantly to the overall fit. Having identified the best model, atypical outliers were identified and removed (employing 2.5 SD of the residual errors as criterion). The models were then refitted to ensure that the results were not driven by a few overly influential outliers. Statistics of the refitted models are reported.

To assess the performance of MSG, we ran the same statistical analysis using as dependent variable MSG-generated passage-based chimera-probe similarity

⁵The full datasets of Experiment 1 and Experiment 2 are available from: <http://clic.cimec.unitn.it/Files/PublicData/chimeras.zip>

	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5	Probe 6
Low inf.	2.65	2.56	2.39	2.31	2.06	2.02
Medium inf.	2.75	2.73	2.41	2.29	2.07	2.07
High inf.	2.87	2.72	2.42	2.27	2.08	2.03
2 sentences	2.71	2.69	2.42	2.28	2.07	2.02
4 sentences	2.74	2.68	2.41	2.28	2.06	2.05
6 sentences	2.83	2.67	2.40	2.31	2.09	2.05

Table 2: Rating distributions in Experiment 2. Average ratings (in different columns) are reported for set of probes grouped and ranked by their ground-truth CPR, crossed with passage informativeness (upper table) and passage length (lower table).

scores in place of human intuitions. The only difference with respect to the analysis of human behavior is that, in the case of model predictions, there are no subjects as aggregation unit, resulting in a simpler random effect structure including only the chimera random intercept. The rationale behind this approach is that a good computational model should not (only) be correlated with human scores, but, more importantly, reproduce the statistical structure subtending behavioral measures: for example, if we observe a particular interaction in the analysis of participants' response, this very same pattern should be observed in the computational simulation predictions (see Baayen et al., 2011, for a similar approach). Still, we also report direct correlation between average human ratings and model predictions.

As an additional control, we further tested the data by including random slopes in the statistical models. These random parameters are meant to capture the variability of a given fixed effect (e.g., CPR) between a series of aggregator units (e.g., subjects), and permit to evaluate whether the considered fixed effect is reliably observed across the considered units. Random slopes of each included predictor were tested in association with chimeras for the analysis on model data, and in association with chimeras and subjects for the analysis on human data. Random slopes were included in the model only if their contribution to the overall model fit was significant. This was tested by a goodness-of-fit test comparing the model before and after the inclusion of the additional random parameter.

3.2. Results

—————Insert Table 2 about here —————
 —————Insert Table 3 about here —————

Table 2 presents a summary of the data collection results for the first experiment. Figure 3 represents the association between the two dependent variables (human ratings and model predictions) with respect to the observed data points. This direct correlation is at a highly significant $r = .39$ ($p = .0001$). Figure 4 represents the association between CPR and human responses (left-hand panel) and model predictions (right-hand panel).

—————Insert Figure 3 about here —————

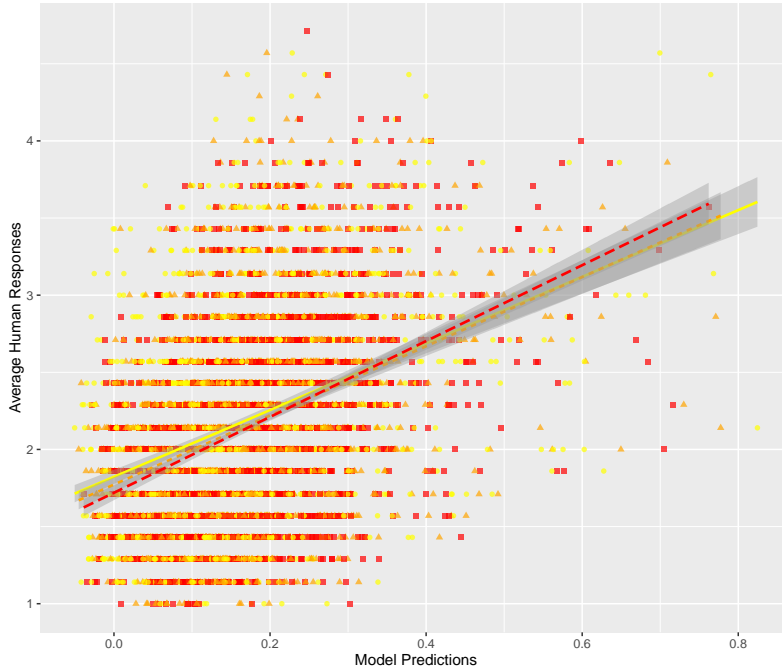


Figure 3: Experiment 1 (word probes): association between by-item average human responses and model predictions. Yellow circles: 2-sentence passages; orange triangles: 4-sentence passage; red squares: 6-sentence passages.

Table 3 and Figure 5 report the results of the mixed-effects analysis on human responses and model predictions. As a result of the outlier-removal procedure, 3.4% of the data points were excluded from the human-rating dataset, and 2.3% of the data points were excluded from the model-prediction dataset.

Subject results are reported in the first three columns of Table 3. We observe a significant interaction between CPR and informativeness: the more informative the presented sentences are, the more markedly subject intuitions match ground-truth CPR. This is also evident from the left panel of Figure 5, representing the interaction captured in the mixed-effects analysis. Different lines in the plot represent the effect of (mean-centered) CPR at different level of informativeness: the higher the informativeness, the steeper the slope, the more aligned human responses are to ground-truth CPR. Moreover, the plot shows how the effect of CPR is evident also for trials in which informativeness is very low. This crucially indicates that participants are able to learn something meaningful from contexts that are not only short, but also poor in explicit evidence. Passage length (that is, the number of sentences forming the context) has no significant modulation on participants' performance (and consequently it is not

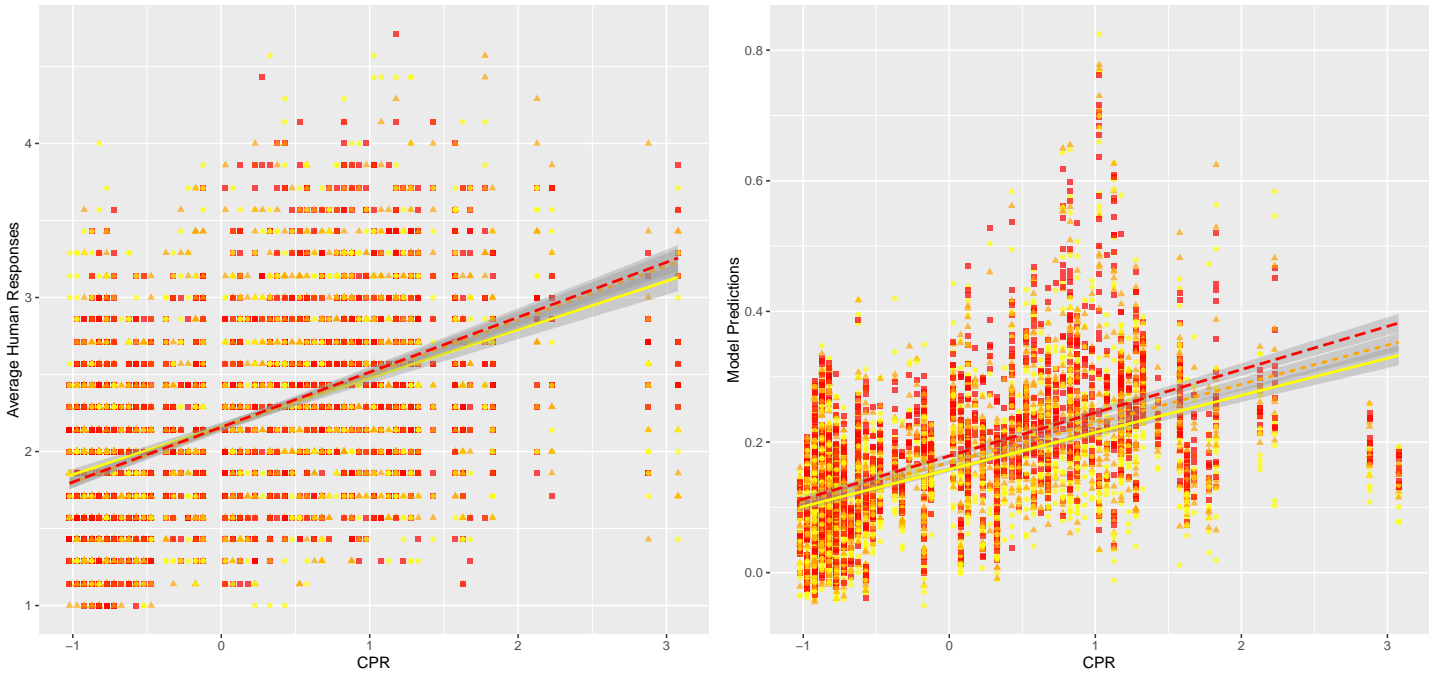


Figure 4: Experiment 1 (word probes): association between CPR and by-item average human responses (left-hand panel) and between CPR and model predictions (right-hand panel). Yellow circles: 2-sentence passages; orange triangles: 4-sentence passage; red squares: 6-sentence passages.

part of the final model): even when only two sentences are presented, the CPR effect does not significantly decrease.

—————Insert Figure 4 about here —————

The next set of columns in Table 3 reports the results when testing the MSG scores for passage-based chimera representations and probes. The pattern is close to what observed for human responses: we found no reliable effect of passage length, and a significant interaction between CPR and informativeness. The latter is also qualitatively very similar to the one found for human participants (compare left and right panel of Figure 5).

—————Insert Figure 5 about here —————

The mixed-effects models significantly improved following the inclusion of CPR random slopes for subjects and chimeras (human rating analysis) and for chimeras (model prediction analysis). However, the reported effects (Table 3) still hold following the inclusion of random slopes, and the pattern of results remains consistent with the one reported in Figure 5.

Predictor	Human responses			Model predictions		
	b	SEM	t	b	SEM	t
Intercept	2.138	0.066	32.44	0.166	0.007	21.87
Informativeness	0.019	0.005	5.73	0.008	0.001	10.49
CPR	0.305	0.003	65.96	0.056	0.001	51.27
Informativeness * CPR	0.031	0.004	8.48	0.008	0.001	9.56

Table 3: Experiment 1 (word probes): Results when analyzing either human responses or model predictions.

4. Experiment 2: Estimating relatedness of chimeric concepts to image probes

4.1. Methods

4.1.1. Participants

As for Experiment 1, participants were recruited through the Crowdfunder platform. Eventually, 168 subjects participated in Experiment 2. As an average, each chimera was rated by 120 distinct subjects.

4.1.2. Procedure

We used the same materials and followed the same procedure as in Experiment 1 above, running 18 separate Crowdfunder jobs on as many lists generated through a Latin-square design. The only difference is that probe words were substituted with images of the corresponding concepts, and ground-truth CPR scores (see Section 2.1.2 above) based on component-word-to-probe-image similarity judgments.

Participants were asked to rate the relatedness between the meaning denoted by the unknown word and the presented image. In the instructions, we stressed the visual aspect of the task, asking participants to base their ratings on their impression, from the passage, of how the unknown concept would look like.

4.1.3. Data Analysis

The data were analyzed as described above for Experiment 1.

4.2. Results

—————Insert Table 4 about here —————

Table 4 presents a summary of the results of the data collection for the second experiment. Figure 6 represents the association between the two dependent variables with respect to the observed data points. The correlation between them is at $r = .34$ ($p = .0001$). Figure 7 represents the association between CPR and human responses (left-hand panel) and model predictions (right-hand panel).

—————Insert Figure 6 about here —————

Table 5 and Figure 8 report the results of the mixed-effects analysis on human responses and model predictions. As a result of the outlier-removal procedure,

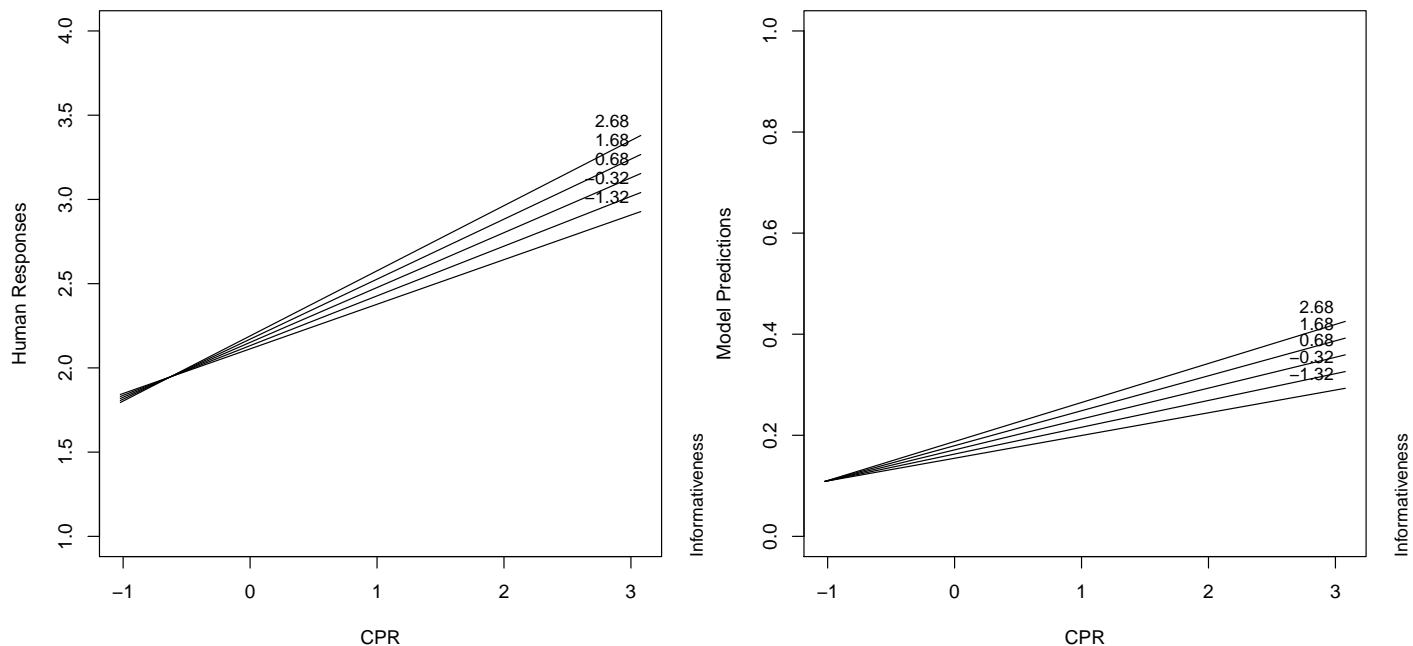


Figure 5: Experiment 1 (word probes): Interaction between informativeness and CPR for human responses (left panel) and model predictions (right panel).

1.6% of the data points were excluded from the human-rating dataset, and 2.1% of the data points were excluded from the model-prediction dataset.

—————Insert Table 5 about here —————

Results parallel Experiment 1. First, we observe significant effects of CPR, informativeness and their mutual interaction. Second, the modulation of passage length did not hold against the statistical controls, and was hence removed from the model. Third, the overall pattern observed in human judgments and model predictions is similar.

Also in this case, the mixed-effects models significantly improved following the inclusion of random slopes. Both CPR and informativeness random slopes (associated to both subjects and chimeras in the human rating analysis, and to chimeras in the model prediction analysis) significantly improved the model fit. Again, this more complex random structure does not change the overall pattern of the reported effects.

—————Insert Figure 7 about here —————

One advantage of a model closely mimicking subject behaviour, such as

	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5	Probe 6
Low inf.	2.52	2.33	2.23	2.12	1.87	1.85
Medium inf.	2.59	2.37	2.19	2.05	1.86	1.82
High inf.	2.71	2.49	2.19	2.01	1.82	1.82
2 sentences	2.59	2.39	2.20	2.08	1.86	1.86
4 sentences	2.60	2.41	2.20	2.02	1.85	1.82
6 sentences	2.65	2.40	2.20	2.06	1.84	1.79

Table 4: Rating distributions in Experiment 2. Average ratings (in different columns) are reported for set of probes grouped and ranked by their ground-truth CPR, crossed with passage informativeness (upper table) and passage length (lower table).

Predictor	Human responses			Model predictions		
	b	SEM	t	b	SEM	t
Intercept	2.245	0.056	39.98	0.045	0.006	7.81
Informativeness	0.014	0.004	3.75	0.004	0.001	6.08
CPR	0.346	0.006	56.93	0.033	0.001	31.05
Informativeness * CPR	0.039	0.005	8.55	0.006	0.001	7.45

Table 5: Experiment 2 (image probes): Results when analyzing either human responses or model predictions.

MSG, is that it can offer insights on the inner workings of word meaning formation. As an example of the possibilities offered by computer simulations, Figure 9 visualizes how MSG “imagines” the thing denoted by a novel word, based solely on the word representation it extracted from a single passage. We stress that what we report here is just an informal analysis of these visualizations, providing preliminary qualitative insight into how distributional concept formation might work, and we leave a systematic study using this methodology to future work.

—Insert Figure 8 about here —

To generate the images in Figure 9, we first obtained a visual vector by mapping the relevant passage-based chimera representation to visual space (using a mapping function from linguistic to visual representations analogous to the inverse one described in Section 2.2.3 above). Then, we retrieved the 50 pictures in our image pool (Section 2.2.2) whose visual feature vectors were nearest to the mapped chimera, and we superimposed them. Intuitively, if we have never seen an object, we might associate it to a mental image that is an average of those objects that we think are similar: Before ever seeing a *papaya*, we might expect it to look somewhere in between a *mango* and a *banana* (see Lazaridou et al., 2015b, for further details on this image generation method).

The first two images in the figure correspond to two-sentence passages about the *caterpillar/cockroach* chimera. The first passage was “easy” for subjects (in the sense that averaged passage-based judgments across probes were highly correlated with ground-truth CPR, $r = 0.95$), and indeed, when deriving its

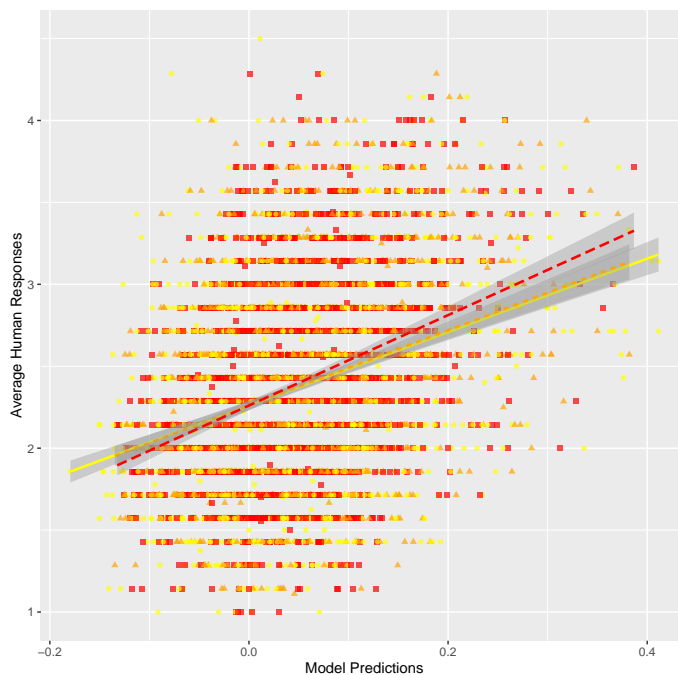


Figure 6: Experiment 2 (image probes): association between by-item average human responses and model predictions. Yellow circles: 2-sentence passages; orange triangles: 4-sentence passage; red squares: 6-sentence passages.

representation of the novel word from this context, MSG visualizes a brownish blob in the middle of green, suggesting a very sketchy bug. The second passage is harder (subjects-CPR correlation: $r=0.46$), and MSG is misled by the metaphorical usage in the second sentence, visualizing a city landscape. The third passage (pertaining to the *potato/turnip* chimera) is relatively easy for subjects ($r=0.71$), as well as for MSG, that is clearly pushed by the first sentence to emphasize the edible aspect of the chimera. The fourth passage was particularly difficult ($r=0.07$), and again we see how MSG is also misled (the image seemingly dominated by *peppermint* and/or *field*).

—Insert Figure 9 about here —

5. Discussion

Despite the fact that distributional learning is typically seen as a long-term process based on large-corpus statistics, our experimental and computational results suggest that this mechanism supports the construction of a reasonable semantic representation as soon as a new word is encountered.

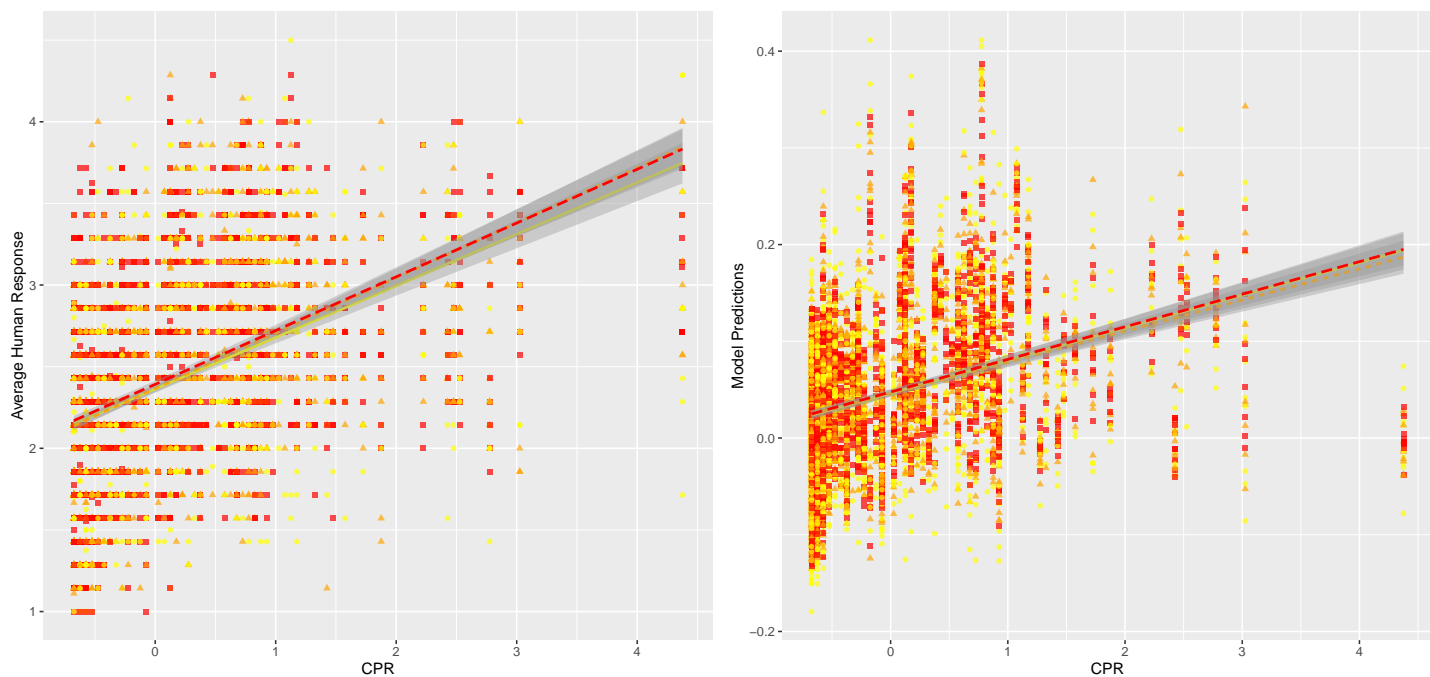


Figure 7: Experiment 2 (image probes): association between CPR and by-item average human responses (left-hand panel) and between CPR and model predictions (right-hand panel). Yellow circles: 2-sentence passages; orange triangles: 4-sentence passage; red squares: 6-sentence passages.

Our experimental data confirmed that purely linguistic distributional evidence constitutes a precious source of information in acquiring word meaning, and that very limited amounts of uncontrolled text, of the sort one naturally encounters in everyday life, suffice for human participants to form reasonably accurate semantic representations, that involve both intuitions about the position of the concepts denoted by the new words in language-based semantic space (Experiment 1), and predictions about their visual aspect (Experiment 2).

Contrary to expectations about evidence accumulation in incremental learning, the length of the passage presented to subjects had no effect on the quality of the representations they extracted. The length difference of our stimuli was limited, ranging from 2 to 6 sentences, and it is likely that more marked differences will significantly impact human performance. Still, the present results indicate that, when evidence is very limited, human learning is mostly determined by the quality of the supplied information (captured here by the informativeness measure), rather than by its quantity.

A version of the MSG distributional semantic model, that we equipped with adult-like semantic competence by pre-training it on natural linguistic and image

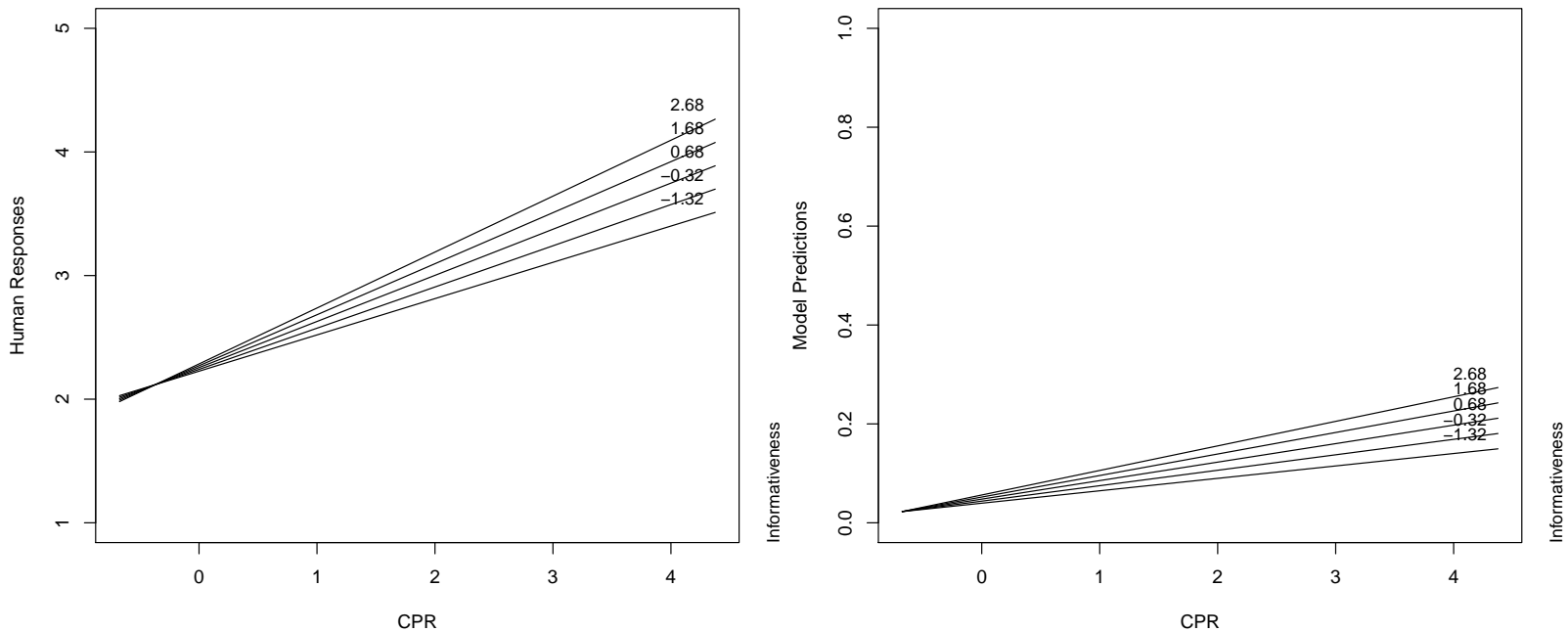


Figure 8: Experiment 2 (image probes): Interaction between informativeness and CPR on human responses (left panel) and model predictions (right panel).

data, performed the word learning tasks in ways that are remarkably close to those of human subjects. Indeed, the reported simulations mirror the interaction between informativeness and CPR observed in our participants, suggesting that the model is not just generating cognitively plausible representations, but also exploiting, in this process, the same type of information that guides human learning.

Our results suggest that the simple context-prediction mechanisms encoded in MSG might suffice to explain the human ability to extract conceptual representations from minimal textual contexts. Specifically, MSG has been trained to produce word representations that are highly predictive of the linguistic contexts in which words occur, and of the visual features of the objects that they denote. Such architecture, when presented with a new word in a passage, can immediately derive a linguistic representation for the word by interpolating over the known words in the context. This produces a representation for the new term that can be used to estimate its degree of relatedness to other words. Moreover, through cross-modal mapping, the model can simulate human intuitions about



- 1) Their food consists of virtually any invertebrate small enough to swallow, including grasshoppers, spiders and **enefies**.
Returning home, they were half way through a portion of lime pickles before finding a 25mm-long **enefy** in the food.
- 2) The **enefies** come in two colour forms - black and a yellow green, both with yellow markings down each flank.
If capitalism can be credited with historic levels of prosperity for many, the inner city is the **enefy** at its heart.
- 3) It started with a green salad, followed by a mixed grill with rice, chips and **scrunts**.
I may be wrong but I don't think **scrunts** would give their farmers the same profit margin as opium poppies.
- 4) If we insert a gene for making oil of peppermint, we'll end up with peppermint flavoured **scrunts**.
The field was resown with ordinary **scrunt**, which only now is starting to make headway.

Figure 9: How MSG visualizes the novel word in each of 4 passages, two constructed from *caterpillar/cockroach* contexts, and two from *potato/turnip* contexts.

how the new concept must look like. In this perspective, the MSG architecture plausibly illustrates how previously acquired knowledge (both linguistically and visually connoted) is exploited to generate novel representations when a new word is first encountered.

The feature-level associationist learning process implemented in MSG can be seen as complementary to the widely studied cross-situational statistics tracking methods needed to link words to their referents (Smith et al., 2014). The MSG learning procedure assumes that the right referent of (a subset of) words is known, and it relies on visual properties of referents (as well as linguistic contexts) to build word meaning representations. It is thus natural to assume that MSG learning operates on input that has been referentially disambiguated through standard cross-situational word-referent linking.

On the other hand, MSG captures aspects of learning that are problematic for classic cross-situational models. For example, it can explain why *similarity* and *context* play an important role in the computation and retention of cross-situational statistics (Vlach and Sandhofer, 2011): similar instances of an object will have similar visual features, speeding up learning. It can explain why word learning, as Smith et al. (2014) put it, is “incremental and slow”, and simply associating a new word to a new referent in an unambiguous context (“fast mapping”) is not guarantee of genuine word learning, that is, full understanding and long-term retention of the word meaning. In the MSG model, the meaning (= vector) of a word is updated at each encounter with the word based on linguistic and visual contexts, as well as indirect influence from other words whose meaning is being constructed in the same representational space. Thus, meaning is indeed acquired incrementally: As our experiments showed, the model can already build a rough sketch of a word meaning from very few exposures, but the representation will keep becoming more precise (leading to better linguistic and cross-modal prediction) as more evidence is seen. Finally, by learning also from purely linguistic contexts, MSG accounts for abstract words, or even

learning about concrete terms in absence of a referent, which remains a mystery for standard cross-situational learning models.

In the future, we intend to pursue the issue of whether word-level cross-situational learning and feature-level MSG-style learning have to be treated as connected but distinct processes, or if they can instead be unified into a single computational architecture. In Lazaridou et al. (2016), we present a preliminary simulation of word learning from child-directed speech suggesting that, indeed, a simple extension of MSG can also account for cross-situational learning, further boosting the idea that DSM-like models might play a central role in explaining human language acquisition.

Acknowledgments

We thank the Cognitive Science editor and reviewers for constructive criticism. We also received useful feedback from the audience at *SEM 2015 and the International Meeting of the Psychonomic Society 2016. We acknowledge ERC 2011 Starting Independent Research Grant nr. 283554 (COMPOSES project). Marco Marelli conducted most of the work reported in this article while employed by the University of Trento. All authors equally contributed to the reported work.

References

- Aitchison, J., 1993. *Words in the Mind*. Blackwell, Malden, MA.
- Baayen, R. H., Davidson, D. J., Bates, D., 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 390–412.
- Baayen, R. H., Milin, P., Durdević, D. F., Hendrix, P., Marelli, M., 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118 (3), 438.
- Baroni, M., Lenci, A., Onnis, L., 2007. ISA meets Lara: An incremental word space model for cognitively plausible simulations of semantic learning. In: *Proceedings of the ACL Workshop on Cognitive Aspects of Computational Language Acquisition*. pp. 49–56.
- Bloom, P., Markson, L., 1998. Capacities underlying word learning. *Trends in Cognitive Sciences* 2 (2), 67–73.
- Borovsky, A., Elman, J., Kutas, M., 2012. Once is enough: N400 indexes semantic integration of novel word meanings from a single exposure in context. *Language Learning and Development* 8 (3), 278–302.
- Borovsky, A., Kutas, M., Elman, J., 2010. Learning to use words: Event related potentials index single-shot contextual word learning. *Cognition* 116 (2), 289–296.

- Deng, J., Dong, W., Socher, R., Li, L.-J., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Proceedings of CVPR. Miami Beach, FL, pp. 248–255.
- Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research* 12, 2121–2159.
- Erk, K., Padó, S., Padó, U., 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics* 36 (4), 723–763.
- Frassinelli, D., Keller, F., 2012. The plausibility of semantic properties generated by a distributional model: Evidence from a visual world experiment. In: Proceedings of CogSci. pp. 1560–1565.
- Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T., 2013. DeViSE: A deep visual-semantic embedding model. In: Proceedings of NIPS. Lake Tahoe, NV, pp. 2121–2129.
- Gillette, J., Gleitman, H., Gleitman, L., Lederer, A., 1999. Human simulations of vocabulary learning. *Cognition* 73 (2), 135–176.
- Howell, S., Jankowicz, D., Becker, S., 2005. A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language* 53, 258–276.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093.
- Keuleers, E., Brysbaert, M., 2010. Wuggy: A multilingual pseudoword generator. *Behavior research methods* 42 (3), 627–633.
- Kievit-Kylar, B., Jones, M., 2011. The Semantic Pictionary project. In: Proceedings of CogSci. Austin, TX, pp. 2229–2234.
- Krizhevsky, A., Sutskever, I., Hinton, G., 2012. ImageNet classification with deep convolutional neural networks. In: Proceedings of NIPS. Lake Tahoe, Nevada, pp. 1097–1105.
- Lake, B., Ullman, T., Tenenbaum, J., Gershman, S., 2016. Building machines that learn and think like people. <https://arxiv.org/abs/1604.00289>.
- Landauer, T., Dumais, S., 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.
- Lazaridou, A., Chrupała, G., Fernandez, R., Baroni, M., 2016. Multimodal semantic learning from child-directed input. In: Proceedings of NAACL. San Diego, CA, In press.

- Lazaridou, A., Pham, N., Baroni, M., 2015a. Combining language and vision with a multimodal skip-gram model. In: Proceedings of NAACL. Denver, CO, pp. 153–163.
- Lazaridou, A., Tien Nguyen, D., Baroni, M., 2015b. Do distributed semantic models dream of electric sheep? Visualizing word representations through image synthesis. In: Proceedings of the EMNLP Vision and Language Workshop. Lisbon, Portugal, pp. 81–86.
- McDonald, S., Ramscar, M., 2001. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In: Proceedings of CogSci. pp. 611–616.
- McRae, K., Cree, G., Seidenberg, M., McNorgan, C., 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods* 37 (4), 547–559.
- Mestres-Missé, A., Rodriguez-Fornells, A., Münte, T., 2007. Watching the brain during meaning acquisition. *Cerebral Cortex* 17 (8), 1858–1866.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS. Lake Tahoe, NV, pp. 3111–3119.
- Miller, G., Charles, W., 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6 (1), 1–28.
- Morin, F., Bengio, Y., 2005. Hierarchical probabilistic neural network language model. In: Proceedings of AISTATS. Barbados, pp. 246–252.
- Nagy, W., Anderson, R., Herman, P., 1987. Learning word meanings from context during normal reading. *American Educational Research Journal* 24, 237–270.
- Padó, S., Lapata, M., 2007. Dependency-based construction of semantic space models. *Computational Linguistics* 33 (2), 161–199.
- Sarri, M., Greenwood, R., Kalra, L., Driver, J., 2011. Prism adaptation does not change the rightward spatial preference bias found with ambiguous stimuli in unilateral neglect. *Cortex* 47 (3), 353–366.
- Schnoebelen, T., Kuperman, V., 2010. Using Amazon Mechanical Turk for linguistic research. *Psihologija* 43 (4), 441–464.
- Schütze, H., 1997. *Ambiguity Resolution in Natural Language Learning*. CSLI, Stanford, CA.
- Smith, L., Suanda, S., Yu, C., 2014. The unrealized promise of infant statistical word-referent learning. *Trends in Cognitive Sciences* 18 (5), 251–258.

- Sternberg, R., Powell, J., 1983. Comprehending verbal comprehension. *American Psychologist* 38, 878–893.
- Trueswell, J., Medina, T., Hafri, A., Gleitman, L., 2013. Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology* 66 (1), 126–156.
- Turney, P., Neuman, Y., Assaf, D., Cohen, Y., 2011. Literal and metaphorical sense identification through concrete and abstract context. In: *Proceedings of EMNLP*. Edinburgh, UK., pp. 680–690.
- Vlach, H., Sandhofer, C., 2011. Developmental differences in children’s context-dependent word learning. *Journal of Experimental Child Psychology* 108 (2), 394–401.
- Werner, H., Kaplan, E., 1950. Development of word meaning through verbal context: An experimental study. *Journal of Psychology* 29, 251–257.
- Weston, J., Bengio, S., Usunier, N., 2010. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning* 81 (1), 21–35.
- Zeiler, M., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: *Proceedings of ECCV (Part 1)*. Zurich, Switzerland, pp. 818–833.