Semantics & Pragmatics Volume 9, Article 17: 1–63, 2016
http://dx.doi.org/10.3765/sp.9.17

What do you know about an alligator when you know the company it keeps? *

Katrin Erk
University of Texas at Austin

Submitted 2015-06-15 / First decision 2015-08-14 / Revision deceived 2016-01-19 / Accepted 2016-02-26 / Final version received 2016-03-02 / Published 2016-04-27

Abstract Distributional models describe the meaning of a word in terms of its observed contexts. They have been very successful in computational linguistics. They have also been suggested as a model for how humans acquire (partial) knowledge about word meanings. But that raises the question of what, exactly, distributional models can learn, and the question of how distributional information would interact with everything else that an agent knows.

For the first question, I build on recent work that indicates that distributional models can in fact distinguish to some extent between semantic relations, and argue that (the right kind of) distributional similarity indicates property overlap. For the second question, I suggest that if an agent does not know what an alligator is but knows that *alligator* is similar to *crocodile*, the agent can probabilistically infer properties of alligators from known properties of crocodiles. Distributional evidence is noisy and partial, so I adopt a probabilistic account of semantic knowledge that can learn from such data.

Keywords: distributional semantics, distributional semantics and formal semantics, language learning

* Research for this paper was supported by the National Science Foundation under grants 0845925 and 1523637. I am grateful to Gemma Boleda and Louise McNally, who read earlier versions of the paper and gave me much appreciated feedback. Many thanks also to Judith Tonhauser and the anonymous reviewers for Semantics & Pragmatics for their extensive and eminently helpful comments. I would also like to thank Nicholas Asher, Marco Baroni, David Beaver, John Beavers, Ann Copestake, Ido Dagan, Aurélie Herbelot, Hans Kamp, Alexander Koller, Alessandro Lenci, Sebastian Löbner, Julian Michael, Ray Mooney, Sebastian Padó, Manfred Pinkal, Stephen Roller, Hinrich Schütze, Jan van Eijck, Leah Velleman, Steve Wechsler, and Roberto Zamparelli for their helpful comments. All remaining errors are of course my own. Thanks also to the Foundations of Semantic Spaces reading group for many discussions that helped shape my thinking on this topic.

1 Introduction

Distributional models characterize the meaning of a word through the contexts in which it has been observed. In the simplest case, these models just record other words that have been observed in the vicinity of a target word in large text corpora, and form some sort of aggregate over the recorded context items. They then estimate the semantic similarity between words based on contextual similarity. In computational linguistics, these simple models have been incredibly successful (Turney & Pantel 2010). They have been used, among other tasks, to find synonyms (Landauer & Dumais 1997) and automatically construct thesauri and other taxonomies (Lin 1998b, Snow, Jurafsky & Ng 2006), to induce word senses from data (Schütze 1998, Lewis & Steedman 2013), to support syntactic parsing (Wu & Schuler 2011), to construct inference rules (Lin & Pantel 2001, Kotlerman et al. 2010, Beltagy et al. 2013), to characterize selectional preferences (S. Padó, U. Padó & Erk 2007), and to aid machine translation (Koehn & Knight 2002).

But are distributional models relevant to semantic theory? This is a question that has been raised in a number of recent papers (Lenci 2008, Copestake & Herbelot 2013, Erk 2013, Baroni, Bernardi & Zamparelli 2014). The most compelling argument for assuming a role for distributional models in semantics is that they provide an explanation of how people can learn something about the meaning of a word by observing it in use (Landauer & Dumais 1997). The argument is that when a speaker repeatedly observes an unknown word in context, they develop an understanding of how to use the word. But what does a speaker know about a word when they know how to use it? Landauer and Dumais write (p. 227):

> Many well-read adults know that Buddha sat long under a banyan tree (whatever that is) and Tahitian natives lived idyllically on breadfruit and poi (whatever those are). More or less correct usage often precedes referential knowledge (E. Levy & Nelson 1994).

This presents a puzzle. If a speaker has no knowledge of the reference of *banyan*, what do they know about the semantics of the word? They clearly know more than nothing, for example they would be able to determine the truth value of the sentence *A banyan tree is a plant*. But they do not know everything about it, for example they would not be able to determine the truth value of the sentence *This is a banyan tree* (spoken in the presence of a

banyan tree). In that case, how can the speaker successfully use the word? What seems to happen is that speakers know that banyan trees are trees, and that breadfruit and poi are food items, so they know some properties of *banyan*, *breadfruit* and *poi* (whose extensions are supersets of the extensions of *banyan*, *breadfruit* and *poi*), and they use the words accordingly. A passage in the famous twin-earth paper of Putnam (1973) raises the same issue as Landauer and Dumais' banyan tree. Putnam writes: "Suppose you are like me and cannot tell an elm from a beech tree. We still say that the extension of 'elm' in my idiolect is the same as the extension of 'elm' in anyone else's, viz., the set of all elm trees" (p.704). So again, if Putnam is not aware of the extension of the word *elm*, then why is he able to use the word felicitously in his article? The answer is the same: If Putnam knows that elms are trees, then he can use the word *elm* accordingly.

The argument that I will make about the banyan tree is that a speaker can successfully use a word in some circumstances by knowing its properties, even when they do not know the word's extension. (I will loosely say "properties of a word" to mean properties that apply to all entities in the word's extension.) I will argue that distributional information can help with inferring a word's properties – and hence, indirectly, some knowledge about the word's extension, as that must be a subset of the extensions of the properties. Suppose I do not know what an alligator is, or more precisely, that I do not know what properties apply to alligators. But I know that an alligator must be something like a crocodile, because it appears in similar textual contexts. I conclude that alligators have many properties in common with crocodiles, so I consider it likely that alligators are dangerous, and also that they are animals. The inferences that can be drawn from the distributional similarity of *alligator* and *crocodile* (called *distributional inferences* below) are uncertain and probabilistic. So this paper will use a probabilistic semantics in order to be able to make use of such probabilistic inferences. This, in a nutshell, is the argument that this paper makes. The paper makes two main contributions. One is to suggest that distributional inference is *property inference*, that is that speakers can probabilistically infer properties based on distributional similarity. The second is a probabilistic inference mechanism for integrating distributional evidence with formal semantics.

Distributional models. What I mean by a distributional model is a mechanism that draws inferences from observed linguistic contexts, in particular from an aggregate of all observed contexts of a target word rather than from

an individual instance (Lenci 2008). (1) shows some sample contexts for the word *alligator* from the British National Corpus.[1]

(1) a. On our last evening, the boatman killed an alligator as it crawled past our camp-fire to go hunting in the reeds beyond.
b. He falls on the floor, belly up, wiggling happily, hands sticking out from the shoulders at a crazy alligator angle.
c. A study done by Edwin Colbert and his colleagues showed that a tiny 50 gramme (1.76 oz) alligator heated up 1 °C every minute and a half from the Sun, while a large alligator some 260 times bigger took seven and a half minutes.
d. The throne was occupied by a pipe-smoking alligator.
e. It was my idea of what an alligator might find appealing.

Sometimes one can learn a lot about a word from a single instance, for example (1a): An alligator is most likely an animal (as it can crawl and can be killed) and a carnivore (as it can go hunting). But not all sentences are like that. There are sentences that are not very informative individually, such as (1c) and (1e), or metaphorical like (1b), and (1d) even describes a fictional world. But by combining weak evidence from these sentences, a distributional model can still derive some information about what an alligator is. This information will necessarily be noisy and probabilistic.

Distributional similarity as indicating property overlap. Until recently it was assumed that distributional models could only estimate "semantic similarity" without being able to distinguish between different semantic relations. That is, *alligator* might come out as similar to *animal*, *crocodile*, and *swamp* – which would make it hard to draw any inferences at all from this evidence. *Animal* is a hypernym of *alligator*, a more general term. *Crocodile* is a co-hyponym of *alligator*, it shares the same direct hypernym (at least it does in some taxonomies). *Swamp* is not related to *alligator*, though this particular non-relation has been jokingly termed "contextonymy", the tendency for two words to be mentioned in the same text passages. It has long been discussed as one of the main drawbacks of distributional models (for example in G. L. Murphy 2002) that if distributional similarity conflates all these semantic relations (or non-relations), no particular inference can be drawn from it.

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

But as it turns out, different types of distributional models differ in what kinds of word pairs receive a high distributional similarity. Distributional models that only count context items close to a target word (*narrow-context* models) tend to rate synonyms, hypernyms and in particular co-hyponyms as similar (Peirsman 2008, Baroni & Lenci 2011), while wide-context models also tend to give high similarity ratings to "contextonyms". In this paper, I argue that narrow-context models do allow for a particular inference, namely one of property overlap: Two words will be similar in such a distributional model if they share many properties, and this happens to be the case with co-hyponyms and synonyms.[2] In this paper I use a broad definition of the term *property* that encompasses hypernyms, and in fact any predicate that applies to all entities in a word's extension.

This raises the question of why it should be possible to draw inferences from a text basis that is as fragmentary and noisy as what we see in (1). An important clue is that only narrow-context models can focus on property overlap to the exclusion of "contextonymy". For noun targets, such narrow contexts will contain modifiers of the target, as well as verbs that take the target as an argument. The noun modifiers often indicate larger categories into which a noun falls. For example, only concrete entities can have colors (if we set aside non-literal uses). Similarly, selectional constraints of verbs indicate semantic properties of the arguments, for example the direct object of *eat* is usually a concrete object and edible (though again non-literal uses as well as polysemous predicates make this inference noisy, but we are not considering them here). If two noun targets agree in many of their modifiers and frequently occur in the same argument positions of the same verbs, then they will tend to share many semantic properties.

The model proposed in this paper. I will argue that while the evidence that comes from distributional models is probabilistic and noisy, that is enough for it to be useful. Even if the agent can only learn that *alligator* and *crocodile* are similar to some degree, that is enough to draw some probabilistic conclusions about alligators. I will use the following three sentences as running examples.

2 Hypernymy, synonymy co-hyponymy, and property overlap are relations between word senses, not words (Fellbaum 1998). I still use the term "relation between words" in this paper, but as I focus on monosemous words only, I use it as a shorthand for the relation between the single senses of two monosemous words.

φ_1: All alligators are dangerous.

φ_2: All alligators are edible.

φ_3: All alligators are animals.

Below I will show a mechanism by which an agent can infer all three sentences based on distributional information. The probability with which an agent distributionally infers a sentence should depend on the strength of the distributional evidence, and by using all three sentences we can test this. Suppose the agent knows things about crocodiles, for example that they are dangerous and that they are animals (but not that they are, in fact, edible). Suppose further that the agent knows things about trouts, for example that they are animals and that they are edible, but the agent knows nothing about alligators. Then sentence φ_1 is an inference that the agent should be able to draw with some certainty from the distributional similarity of *alligator* and *crocodile*. The agent should also ascribe some likelihood to φ_2. But as *crocodile* is more distributionally similar to *alligator* than *trout* is, the agent should be more certain about φ_1 than φ_2. Sentence φ_3 is a conclusion that the agent can draw from a distributional comparison of *alligator* to either *crocodile* or *trout*. In fact, as this conclusion is supported by two pieces of distributional evidence, both *alligator/crocodile* and *alligator/trout*, the agent should be more certain about φ_3 than either φ_1 or φ_2.

As I have argued above, the inferences that arise from distributional evidence should be modeled as probabilistic because this evidence is noisy. This paper uses probabilistic logic (Nilsson 1986), which defines a probability distribution over worlds, to describe an agent's knowledge as a *probabilistic information state*. The probability of a world is the probability that the agent ascribes to that world being the actual world. So suppose again that the agent does not know what an alligator is, but observes a high distributional similarity for *alligator* and *crocodile*. Then this should be reflected in the agent's probabilistic information state. Based on the high distributional similarity, the agent should ascribe higher probability to worlds in which alligators are dangerous and animals (assuming that those are crocodile properties) than to worlds in which that is not the case. And when worlds in which all alligators are dangerous tend to have higher probability than worlds where some alligators are harmless, then the agent will ascribe a higher probability to the sentence "all alligators are dangerous".

Questions not handled. This paper takes a first step in the direction of integrating formal and distributional semantics in a probabilistic inference framework. I will make some simplifying assumptions to keep the task manageable. I focus on distributional learning of properties for monosemous noun concepts only, as nouns are generally easier to characterize in terms of properties than other parts of speech. I will also assume that each target has only a single sense. So while (1) includes metaphoric uses to show the breadth of distributional data "in the wild", I do not handle metaphoric uses in this paper.

I ignore many important questions. In terms of distributional learning, I do not consider the task of learning properties that do not pertain to all members of an extension, or learning properties of polysemous words. I also do not look into other ways, besides learning word meaning, in which distributional information may be relevant, such as determining what a polysemous word means in a given context. This paper is also is preliminary in terms of the probabilistic framework it uses, which can currently only handle finite sets of worlds. Still I believe that the current proposal can serve as a first step in exploring the connection of formal and distributional semantics.

Plan of the paper. Section 2 introduces distributional models and their parameters, as well as the particular distributional model that will be used for examples throughout the paper. Section 3 relates findings from the literature that indicate that not all distributional models have the same notion of "similarity", crucially narrow-context and wide-context models differ in the types of word pairs they judge similar. Section 4 then addresses the first of the two core points of the paper. In this section I argue that what narrow-context models are actually measuring is similarity in terms of properties. Section 5 specifies the probabilistic logic of Nilsson 1986 and uses it to define probabilistic information states. Section 6 addresses the second core point of the paper: a mechanism for probabilistic inference from distributional data. It shows how distributional evidence can probabilistically influence an agent's probabilistic information state. Section 7 revisits the three sentences from above (all alligators are dangerous/edible/animals) to test what probabilities they are assigned by the probabilistic inference mechanism. This is followed by a section that sketches some other approaches that aim to link formal semantics and distributional information (Section 8).

The *boatman killed an* alligator *as it crawled* past.
Snake eats alligator *by swallowing it* whole.
. . . been *eaten by an* alligator.
Alligators *eat fish, birds* and mammals

a-DT	as-IN	bird-NN	boatman-NN	by-IN	crawl-VB	eat-VB	fish-NN
2	1	1	1	2	1	3	1

it-PP	kill-VB	snake-NN	swallow-VB
2	1	1	1

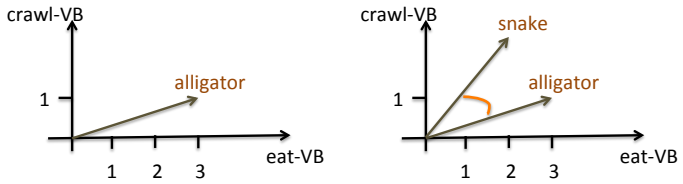


Figure 1 A toy distributional model computed from 4 sentences, 3-word context window (italics), lemmatized and tagged with part of speech. Target word (underlined) is *alligator*. In the middle: table of counts. Bottom left: vector interpretation of the co-occurrence counts, dimensions *eat-VB* and *crawl-VB*. Bottom right: an illustration of the computation of cosine similarity.

2 Distributional models

This section introduces distributional models informally through some toy examples and then defines them formally. It also gives the specifications of the distributional model that will be used for experiments throughout the paper.

2.1 An introduction to distributional models through toy examples

In building a distributional model, we have a number of choices, also called *parameters* of the model. I introduce them by way of example, showing each parameter in italics as it is introduced.

A distributional model provides representations for particular *target words* by counting *context items* in a *corpus*, as illustrated in the top panel of Figure 1. This particular corpus consists of occurrences of the single target word *alligator*. Around each occurrence of *alligator*, we count context items, in this case words. We could have counted the observed word forms, but to illustrate a different choice, we count their lemmas combined with their part-of-speech tag. So for example we count *kill-VB* instead of the observed form *killed*, as shown in the resulting table of counts in the middle of Figure 1. The tags used here are DT for determiner, IN for preposition, JJ for adjective, NN for noun, and VB for verb.

Context items are counted only if they appear close to the target word, that is, if they are within the *relevant context*. Here the relevant context is defined as a three-word window on either side of the target word, not crossing sentence boundaries. There are also distributional models that define the relevant context to comprise the sentence or even the whole document in which the target occurs.

Instead of counting words in a context window around the target, we could also have used syntax to define the relevant context, and could have counted "parse snippets" rather than words as context items. For example, if *water* is the target in the dependency parse in Figure 2, then *mod_muddy/JJ* and *mod-in*$^{-1}$*_like/VB* are context items that are counted: They are in the relevant context, which is defined as consisting of all context items that directly border on the target node in the parse tree. (The $^{-1}$ here is to signal that *water* is the dependent, not the head, of *like*.)[3]

As illustrated on the bottom left of Figure 1, the counts for *alligator* can be interpreted as a vector in a high-dimensional space whose dimensions (also called features) correspond to context items. The dimensions in the illustration are *eat-VB* and *crawl-VB*. The illustration is only showing two dimensions because more would be hard to draw. The actual space encompasses a dimension for each of the context items in the table of counts.

A central use of distributional models is to the predict semantic similarity of two words based on their distributional similarity. There are many possible *similarity measures* that can be used to compute the distributional similarity of two words. Many of them are based on the interpretation of counts as

3 Not all distributional models take the context items to be linguistic. A number of recent models have explored a combination of textual and perceptual context items (Andrews, Vigliocco & Vinson 2009, Feng & Lapata 2010, Bruni et al. 2012). I do not pursue this option in this paper and consider only distributional models that are text-based.

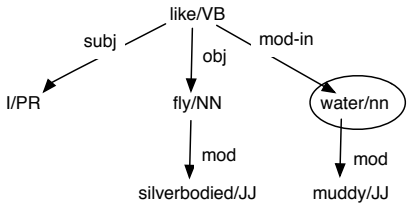


Figure 2 A dependency parse of the sentence "In muddy water I like silverbodied flies", produced by the C&C parser (Curran, Clark & Bos 2007). Syntactic neighbors of the target *water* are *mod_muddy/JJ* and *mod-in*$^{-1}$*_like/VB*.

vectors in a high-dimensional space. The most widely used similarity measure is the cosine of the angle between two vectors, illustrated on the bottom right in Figure 1. In general, say the distributional representation of a word u is $\vec{u} = \langle u_1, \ldots, u_n \rangle$, a vector of real numbers, and likewise the distributional representation of v is $\vec{v} = \langle v_1, \ldots, v_n \rangle$. Then their cosine is the dot product of the two vectors (the sum of the component-wise product of the vectors), normalized by the product of vector lengths:

$$\cos(\vec{u}, \vec{v}) = \frac{\sum_i u_i v_i}{||\vec{u}||\,||\vec{v}||} \tag{1}$$

The length of the vector $\vec{u}$ is the square root of the dot product of the vector with itself, $||\vec{u}|| = \sqrt{\sum_i u_i^2}$. For a concrete example, the representation of the word *alligator* is $\overrightarrow{\text{alligator}} = \langle 3, 1 \rangle$ if we use only the two dimensions illustrated in the two bottom panels of Figure 1. Now say the representation of the word *snake* is $\overrightarrow{\text{snake}} = \langle 2, 3 \rangle$. Then their cosine similarity is

$$\cos(\overrightarrow{\text{alligator}}, \overrightarrow{\text{snake}}) = \frac{3 \cdot 2 + 1 \cdot 3}{\sqrt{3^2 + 1^2}\sqrt{2^2 + 3^2}} = \frac{9}{\sqrt{10}\sqrt{13}} = 0.79$$

For an example of a less similar word, suppose the representation of *skin* is $\overrightarrow{\text{skin}} = \langle 0, 1 \rangle$. Then the cosine similarity of *alligator* and *skin* is

$$\cos(\overrightarrow{\text{alligator}}, \overrightarrow{\text{skin}}) = \frac{0 + 1}{\sqrt{10}\sqrt{1}} = 0.32$$

For the remaining parameters, we switch to a different toy model, shown in Table 1. The top left panel shows representations for four target words, *apple*, *street*, *pass*, and *truck*, with co-occurrence counts (from a hypothetical

	crab	car	grass	tree	orange
apple	3	0	2	5	3
street	0	1	2	7	0
pass	2	5	1	0	5
truck	0	1	0	1	1

	crab	car	grass	tree	orange
apple	0.59	0.0	0.18	0.14	0.0
street	0.0	0.0	0.44	0.74	0.0
pass	0.18	0.76	0.0	0.0	0.51
truck	0.0	0.62	0.0	0.0	0.37

	d_1	d_2	d_3
apple	0.11	0.36	0.51
street	0.03	0.83	-0.22
pass	0.93	-0.03	0.03
truck	0.71	-0.06	-0.11

Table 1 Transforming a distributional space: Table of (made-up) observed counts (top left), its positive point-wise mutual information (PPMI) transformation (bottom left), and singular value decomposition (SVD) transformation of the PPMI table (right)

corpus) for five context items, *crab*, *car*, *grass*, *tree*, and *orange*. These counts (sometimes called "raw counts") can optionally be transformed into *association weights*. Some words, in particular function words like *of* or high-frequency content words such as *say*, will co-occur frequently with all targets. But knowing that all targets co-occur with *of* does not tell us much about how target words differ in their meanings. What we want to know is which context words are most associated with particular targets: which context words appear frequently with some targets but not others. There are several methods for transforming raw counts into association weights. One popular choice is *point-wise mutual information (PMI)*. The point-wise mutual information association weight of a target word v and a dimension (feature) d of the space is the logarithm of the ratio between the observed co-occurrence probability $P(v,d)$ and the expected co-occurrence probability if there is no association, $P(v)P(d)$. The latter is the probability of co-occurrence if the occurrence of v and the occurrence of d are statistically independent.

$$\text{PMI}(v,d) = \log \frac{P(v,d)}{P(v)P(d)} \tag{3}$$

Positive point-wise mutual information (PPMI) is point-wise mutual association if it is positive, and zero otherwise.

$$\text{PPMI}(v,d) = \begin{cases} \text{PMI}(v,d) & \text{if greater zero} \\ 0 & \text{else} \end{cases} \tag{4}$$

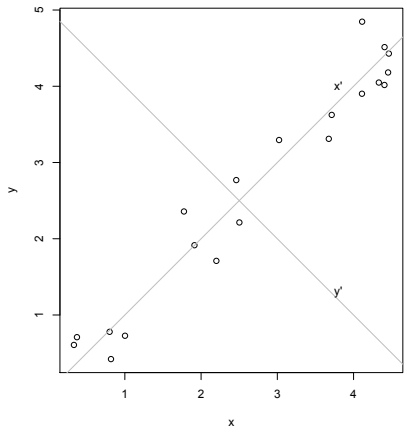


Figure 3 Illustrating dimensionality reduction: This data can equivalently be represented through the dimensions indicated by the gray lines. These are the direction of most variance in the data (x') and the direction of most remaining variance in the data (y')

All relevant probabilities can be computed from the table of raw counts. The co-occurrence probability $P(v,d)$ of the target v and feature d is the relative frequency of the target/feature pair, $P(v,d) = \frac{\#(v,d)}{\#(\_,\_)}$. We write $\#(v,d)$ for the co-occurrence count of v with d, and $\#(\_,\_)$ for the summed co-occurrence count of any target with any feature (so it is the sum of all counts in the table). The probability of the target v is its relative frequency $P(v) = \frac{\#(v,\_)}{\#(\_,\_)}$, where $\#(v,\_)$ is the summed co-occurrence count of target v across all features. The probability of the feature d is its relative frequency $P(d) = \frac{\#(\_,d)}{\#(\_,\_)}$, where $\#(\_,d)$ is the summed co-occurrence count of feature d across all targets. PPMI transformation is illustrated in the left two panels of Table 1: The table in the upper left panel shows (hypothetical) raw counts, and the table in the lower left panel is its PPMI transformation. For example, the entry for *apple* and *crab* is $\log \frac{\frac{3}{39}}{\frac{13}{39}\frac{5}{39}} = \log 1.8 = 0.59$.

The distributional representations can optionally be further transformed by *dimensionality reduction*. This technique reduces the number of dimensions of a model, thus making it more manageable. The new dimensions that it introduces can be seen as latent semantic classes (Landauer & Dumais 1997). A popular choice of dimensionality reduction method is singular value decomposition (SVD). SVD involves representing a set of points in a differ-

ent space (that is, through a new set of dimensions) in such a way that it brings out the underlying structure of the data. This is illustrated in Figure 3. The data points in this figure can be described through coordinates on dimensions x and y. Alternatively, the data points can be described through coordinates on x' and y'. These two lines are not chosen at random: Line x' is the direction of most of the variance in the data. Line y' is the direction of all the remaining variance in the data once x' has been accounted for. More generally, SVD will choose new dimensions that are ordered by the amount of variance in the data that they explain. The advantage of this method becomes apparent when we have more than two dimensions. If, say, we have 10,000 dimensions, and the first 300 dimensions in the "new space" describe 99% of the variance in the data, we can safely omit the remaining 9,700 dimensions without losing much information. I will not go into details, but briefly, SVD is a method that factorizes an arbitrary matrix A into three matrices U, Σ, and V that, multiplied together, reconstitute A (where V^T is V with rows and columns transposed):

$$A = U \Sigma V^T \tag{5}$$

Crucially, if each row of A is the distributional representation of one target word, then the rows of $U\Sigma$ represent the same targets in a different set of dimensions, where the dimensions (columns of $U\Sigma$) are ordered by the amount of variance in the data that they explain. So by only using the first k columns of $U\Sigma$, we have reduced the dimensionality of our representation. The right table in Table 1 shows the representations of the target words in a reduced space of three dimensions. The dimensions are now nameless: While the original dimensions stood for individual context items, the new automatically generated dimensions do not. These new dimensions can be viewed as combining observed context items (old dimensions) into latent semantic classes (new dimensions), fuzzy groupings of context items that have similar co-occurrence patterns with targets.

2.2 A formal definition of distributional models

We now formally define distributional models as tuples of their relevant parameters. Most parameters have been introduced informally above: a set T_D of target words that receive distributional representations, a set O_D of context items to be counted in a corpus C, a similarity measure S_D, a choice

of relevant context in which to look for context items, and the options to compute association weights and to do dimensionality reduction.

$T_\mathcal{D}$ and $O_\mathcal{D}$ are arbitrary sets. We add a third set, the set of *basis elements* $B_\mathcal{D}$ that label the dimensions in the space that is eventually constructed. This can be the same as $O_\mathcal{D}$ (as in the two tables on the left of Table 1), but if dimensionality reduction is used the basis elements will not be context items: In the right panel of Table 1 the basis elements are d_1, d_2, d_3. We define a corpus as a collection of targets and context items, $C \in (O_\mathcal{D} \cup T_\mathcal{D})^*$. The similarity function, which maps pairs of targets to a value indicating a degree of distributional similarity, has the signature $S_\mathcal{D} : (T_\mathcal{D} \times T_\mathcal{D}) \rightarrow \mathbb{R}$. We describe the relevant context as an *extraction function* $X_\mathcal{D}$ that takes as input the whole corpus C and returns co-occurrence counts of all targets with all observable items from the set $O_\mathcal{D}$– that is, it returns a mapping from target/context item pairs to numbers in $\mathbb{N}_0$ (as counts can also be zero). In sum, it has the signature $X_\mathcal{D} : (T_\mathcal{D} \cup O_\mathcal{D})^* \rightarrow ((T_\mathcal{D} \times O_\mathcal{D}) \rightarrow \mathbb{N}_0)$. We lump the optional association weight computation and dimensionality reduction steps into a single parameter, an aggregation function $A_\mathcal{D}$ that takes the output of $X_\mathcal{D}$ and turns it into a mapping from targets and basis elements to real values, $A_\mathcal{D} : ((T_\mathcal{D} \times O_\mathcal{D}) \rightarrow \mathbb{N}_0) \rightarrow ((T_\mathcal{D} \times B_\mathcal{D}) \rightarrow \mathbb{R})$.

In summary, we describe distributional models as tuples comprising the sets of target elements, observable items, basis elements, the corpus, the extraction and aggregation function, and the similarity function:

$$\mathcal{D} = \langle T_\mathcal{D}, O_\mathcal{D}, B_\mathcal{D}, C, X_\mathcal{D}, A_\mathcal{D}, S_\mathcal{D} \rangle \tag{6}$$

The term "distributional" has been used for a number of different approaches in computational linguistics. The definition above fits what Baroni, Dinu & Kruszewski (2014) call "count-based" models, which uses counts of co-occurring context items that are compiled into vectors in a high-dimensional space (Turney & Pantel 2010). A second class of approaches that is getting increasingly popular uses neural networks to compute word vectors (then called embeddings) based on some prediction task (Collobert & Weston 2008, Mikolov et al. 2010); some of these approaches will be covered by the definition above, but maybe not all. Bayesian topic models (Blei, Ng & Jordan 2003) constitute a third class of distributional models. They are not covered by the definition above but could be covered by a slight variant.

2.3 The influence of parameters on distributional models

The choice of parameters has a large influence on the types of predictions that a distributional model will make. This is most obvious in the choice of corpus C: Larger corpora are better in general as they yield more stable estimates of co-occurrence frequency. Another way in which the choice of corpus influences the model is through the genres that are represented in it. Lin (1998a), using a distributional model built from newspaper text, found that the words *captive* and *westerner* were respective nearest neighbors, that is, *westerner* was the word most similar to *captive* among all the target words, and vice versa, a finding that clearly derives from kidnapping stories in the newspapers. Note that such genre effects do not mean that the distributional model is useless, just that it is noisy. Today most distributional approaches choose to use the largest amount of corpus data possible to achieve the most stable estimates while diluting any genre effects.

As mentioned in the introduction, the extraction function X_D is an important parameter that influences what a high similarity value means. Narrow-context models tend to give high ratings preferably to word pairs that are co-hyponyms (*alligator/crocodile*) or synonyms, while wide-context models also give high ratings to pairs like *alligator/swamp* that stand in no traditional semantic relation but are topically related. I discuss this parameter in more depth in the following section.

Another parameter that influences what kinds of word pairs are given high similarity ratings is the similarity function S_D itself. Cosine tends to give high ratings to co-hyponym pairs, while other more recent measures aim to give high ratings to hyponym/hypernym pairs (Lenci & Benotto 2012, Roller, Erk & Boleda 2014, Fu et al. 2014).

In general, using association weights instead of raw counts leads to a large improvement in performance. For dimensionality reduction the case is not that clear. It sometimes leads to better performance, but not always. But a model with 300 dimensions is usually better manageable than one with 10,000.

2.4 The distributional model used in this paper

All examples of distributional data in this paper use a common distributional model. It is based on a corpus C that is a concatenation of the English Gigaword Corpus (news), the British National Corpus (mixed genres), a Wikipedia

dump called the English Wackypedia (encyclopedia text), and the ukWaC corpus (web text) (Baroni, Bernardini, et al. 2009). The corpus was automatically lemmatized and tagged with parts of speech. Only nouns, proper nouns, adjectives and verbs with a frequency of at least 500 were retained. The resulting corpus has roughly 2.8 billion tokens. As targets we use all lemma/tag pairs in this corpus; as context items we use the 20,000 most frequent lemma/tag pairs. Occurrences of context items are counted if they are in a window of 2 words on either side of the target, not crossing sentence boundaries.

The aggregation function A_D in this model transforms counts to association weights using PPMI (Equation (4)), and reduces the space to 500 dimensions using SVD. We use cosine (Equation (1)) to compute distributional similarity.[4]

3 Semantic relations: similarity versus relatedness

Having introduced distributional models and their parameters in general in the previous section, I would now like to focus on one particular parameter, the extraction function X_D. As mentioned above, distributional models have often been criticized (for example in G. L. Murphy 2002) for only yielding vague "semantic similarity" ratings that do not provide evidence for any particular semantic relation. But two recent studies indicate that with some choices of the extraction function X_D, word pairs with high similarity ratings will typically be related by a specific semantic relation, or a specific group of relations. This section discusses these studies in detail (so its aim is not to break new ground, but to report relevant results from the literature).

The question of whether distributional models can distinguish between semantic relations is critical for this paper because it is also a question of inferences. If distributional models cannot distinguish between the relations of *alligator/crocodile* and *alligator/jaw*, then it would be hard to see what could be inferred from *alligator* being similar to *crocodile*. But if it is possible to build a distributional model in which pairs like *alligator/crocodile* are

4 When distributional models are used to model human language learning, any automatic processing like lemmatization or part-of-speech tagging is often avoided because this information is not present in the input that humans receive. In this paper, I do not try to model infant language learning, where this is a clear concern. Instead, I model probabilistic inference by a competent speaker about individual unknown words. We can therefore assume that the speaker is familiar with the lemmas and parts of speech of most of the context items observed to co-occur with the target words.

judged as similar, but pairs like *alligator/jaw* are not, then we can draw inferences about *alligator* from a high similarity rating to *crocodile*. (Which inferences exactly we can get from a high similarity to *crocodile* is the subject of Section 4.) Note that if distributional models can differ in this way, then the term "similarity function" is actually somewhat misleading: It suggests that all similarity functions approximate the same notion of similarity, while in fact the function $S_{\mathcal{D}}$ gives ratings that need to be interpreted differently based on the definition of the function itself and based on the other parameters of the model, in particular the extraction function $X_{\mathcal{D}}$.

It will be helpful to have terminology that distinguishes between pairs like *alligator/crocodile* on the one hand and *alligator/jaw* on the other. We take over the distinction made by Peirsman 2008 and Agirre et al. 2009. They both distinguish between *similarity* and *relatedness*. A pair of words is called similar if they are synonyms, hypernym and hyponym, or co-hyponyms. Two words are called related if they do not stand in a synonymy, hypernymy, or co-hyponymy relation but are still topically connected, such as *alligator/jaw* or *plane/pilot*. I will use the terms *AP-similarity* and *AP-relatedness* to signal these specific definitions of similarity and relatedness ("AP" for Agirre and Peirsman).

It has long been known anecdotally that the way to get high similarity ratings for AP-similar words is to use a narrow context window of maybe two or three words on either side of the target word, or a syntactically defined context as illustrated in Figure 2, see for example Sahlgren 2006. More recently, there have been two studies with empirical tests on the context window effect: Peirsman 2008 for Dutch and Baroni & Lenci 2011 for English.

In the introduction I have already briefly discussed the question of why the choice of context window can have such an effect: It is because with a narrow context window, the context items found for a noun target will often be modifiers, or verbs that take the target as an argument. In either case, the context items indicate selectional constraints that the noun target meets – and such constraints are typically expressed in terms of properties. If two noun targets agree in many of the modifiers or verbs that they appear with, then they will typically share many semantic properties, like synonyms or co-hyponyms do. In a wide context window, the context items will be much more mixed.

Back to the two empirical studies of the context window effect. Peirsman (2008) tests several distributional models trained on Dutch newspaper text, using cosine to compute similarity and varying the context window between

concept	*concept class*	*relation*	*relatum*
alligator-n	amphibian_reptile	cohypo	toad-n
alligator-n	amphibian_reptile	hyper	carnivore-n
alligator-n	amphibian_reptile	mero	jaw-n
alligator-n	amphibian_reptile	attri	frightening-j
alligator-n	amphibian_reptile	event	bask-v
alligator-n	amphibian_reptile	random-j	constructive-j
alligator-n	amphibian_reptile	random-n	trombone-n
alligator-n	amphibian_reptile	random-v	fetch-v

Table 2 Sample entries from the *BLESS* dataset

1 and 20 words (ignoring sentence boundaries). He also tests a distributional model where context is defined as syntactic neighborhood (as in Figure 2), and similarity is again computed using cosine. Peirsman tests how many of the target nouns are synonyms, co-hyponyms, or direct hypo- or hypernyms of their single nearest distributional neighbor (that is, the noun that has the highest similarity rating to the target). Peirsman finds that the proportion of nearest neighbors that are AP-similar is largest for the syntactic-context distributional model, and that for the context-window models it generally decreases with window size. This study allows for a tentative conclusion that models with a syntactic definition of context or with a narrow context window have a tendency to give high similarity ratings to AP-similar words. The study also finds that across distributional models, the largest percentage of AP-similar nearest neighbors tend to be co-hyponyms.

For English, the *BLESS* dataset of Baroni & Lenci 2011 was designed to compare the similarity values that a distributional model assigns to words in different semantic relations. Table 2 shows some example entries from the dataset. *BLESS* has pairs of a *concept* and *relatum*, where a concept is an unambiguous concrete noun. The pairs are classified into one of eight semantic relations: *cohypo* is co-hyponymy, *hyper* is hypernymy, *mero* is meronymy, *attri* and *event* are typical attributes and events related to the concept. In addition, there are three random relations for different parts of speech, where *random-n* pairs a concept with a random noun, *random-v* pairs it with a random verb, and *random-j* with a random adjective. The dataset also characterizes each concept into a larger concept class.

Baroni & Lenci 2011 use the *BLESS* dataset to compare the similarity ratings that different distributional models assign to word pairs in particular

semantic relations. They evaluate distributional models in which all parameters are kept fixed except for the context size: two models that use context windows, one narrow (2 content words on either side of the target word) and one wide (20 content words either side), and a model that considers the whole document as the context for an occurrence of the target word. The similarity measure in these experiments is again cosine. To evaluate the models, they determine the similarity of each target to its nearest (most similar) relatum in each of the 8 relations. They then compare the collected similarities for co-hyponyms, hypernyms, meronyms, and so on. In all three distributional models, co-hyponyms have the highest similarity to the targets. But in the document-based model, the difference of co-hyponyms particularly to hypernyms, meronyms, and events is very slight. It is more marked in the 20-word context window model, and quite strong in the 2-word context window model. That is, a document-based model captures a wide variety of relations with similar strength, while a narrow context window makes co-hyponymy prominent. I repeated the experiment with a distributional model with syntax-based context and found that, as expected, models with a syntactic context show similar results to narrow context window models.

4 Property Overlap

In both experiments reported in the previous section, Peirsman 2008 and Baroni & Lenci 2011, narrow context window and syntactic context window models tend to give the highest similarity ratings to AP-similar word pairs, in particular co-hyponyms. In this section we ask which inferences an agent can draw from observing that two words are highly AP-similar. This would be easy if all AP-similar word pairs were in a single semantic relation, but AP-similarity encompasses synonymy, hypernymy, and co-hyponymy.

One could argue that as distributional models tend to give the highest AP-similarity ratings to co-hyponym pairs, we should take AP-similarity simply as an indication of co-hyponymy and ignore the other relations. But co-hyponymy is not well-defined. Because it is the relation between sister terms in a taxonomy, it depends strongly on details of the taxonomy. In *BLESS*, where the relevant higher-level concept class for *alligator* is *amphibian_reptile*, its co-hyponyms are *crocodile*, *frog*, *lizard*, *snake*, *toad* and *turtle*. In WordNet (Fellbaum 1998), a more fine-grained taxonomy, the direct hypernym of *alligator* is *crocodilian*. There *crocodile* is still a sister term of *alligator*, but *frog* and the others are not.

Instead I am going to argue that AP-similarity can be characterized as property overlap, where "property" is again meant in a broad sense that encompasses hypernyms and in fact any predicates that apply to all members of a category. Co-hyponyms have many properties in common. This includes their joint hypernyms: Alligators and crocodiles are both reptiles, and they are both animals. They also share other properties: Alligators and crocodiles are both green, both dangerous, and both scaly. Property overlap also accommodates relations other than co-hyponymy that are included in AP-similarity: Synonyms share most of their properties, and likewise hypernym-hyponym pairs. On the other hand, words that are only AP-related, like *alligator/swamp*, do not usually share many properties. Interpreting AP-similarity as property overlap allows us to draw inferences from observed distributional similarity: If *alligator* and *crocodile* are distributionally highly similar, we infer that they must have many properties in common.

But before we conclude that AP-similarity is property overlap, we need to consider some possible counter-arguments. First, given that co-hyponyms generally receive the highest similarity ratings from distributional models of AP-similarity, we need to ask if property overlap is the best possible characterization of co-hyponymy. Typical examples of co-hyponymy are *alligator* and *crocodile*, or *cat* and *dog*. These pairs share many properties – but they are also incompatible, that is, there is no entity that is both an alligator and a crocodile, or both a cat and a dog. If all co-hyponym pairs were incompatible, then we might have to characterize AP-similarity as something like property overlap plus incompatibility. But as Cruse (2000) points out (Section 9.1), many co-hyponyms in existing taxonomies are not incompatible. If we look at the direct hyponyms of the word *man* (in the sense of *adult male*) in WordNet, we find terms like *bachelor*, *boyfriend*, *dandy* and *father-figure*, which are certainly compatible. In *BLESS*, the larger category *building* contains *castle*, *hospital*, *hotel*, *restaurant*, *villa*. Some of these words are probably incompatible, but not all. So as co-hyponymy does not entail either compatibility or incompatibility, we can leave compatibility out of the picture when interpreting AP-similarity.

Another possible counter-argument to interpreting AP-similarity as property overlap is that relatively few hypernym-hyponym pairs get high AP-similarity ratings, even though hyponyms and hypernyms overlap in all the properties of the hypernym. To see why few hypernym-hyponym pairs get high AP-similarity ratings, it is useful again to look at what distributional models do. Distributional similarity is high if two target words occur in many

of the same contexts, which, as I have argued above, is the case if they share many properties – but this is not the only factor that determines whether two words will occur in the same contexts. Similar level of generality and similar register are two other relevant factors. The words *dog* and *mammal* will rarely occur in the same contexts because *mammal* is a much more formal term than *dog*, as well as a much more general concept. What follows from that is that we should take high distributional similarity as a sign of high property overlap, but we should not take low distributional similarity as a sign of low property overlap. Rather, low distributional similarity should just be read as a sign that there is too little information to judge property overlap. The model that I formulate in Section 6 will do exactly that: It will draw inferences from high degrees of distributional similarity, but not from low distributional similarity.

Existing work on inferring properties from distributional data. There is some empirical evidence that distributional data can be used for inferring properties in Johns & Jones 2012, Făgărăşan, Vecchi & Clark 2015, Gupta et al. 2015, and Herbelot & Vecchi 2015. They test whether distributional vectors can be used to predict a word's properties (where, as above, I use the term "properties of a word" to mean properties that apply to all entities in the word's extension). To do so, they either make use of distributional similarity directly, or use regression to learn a mapping from distributional vectors to "property vectors". All except Gupta et al. use as data the feature norms in McRae et al. 2005 and in Vigliocco et al. 2004. This is data collected through experiments in which participants were asked to provide definitional features for particular words. I take these papers as preliminary evidence that distributional data can indeed be used as a basis for property inference. However, I cannot use any of these approaches directly in this paper; some compute (uninterpretable) weights rather than probabilities, some have a different task from the one I am addressing.[5]

There are also several recent papers that focus specifically on the prediction of hypernyms from distributional data (Lenci & Benotto 2012, Roller,

5 The most similar approach to mine is the one of Herbelot and Vecchi. They have a set-theoretic motivation for the property vectors they use. The numbers in their property vectors encode whether "all", "most", "some", or "few" members of a category have a particular property, where each quantifier is represented by a fixed probability. But the approach does not fit into the setting I use because it is unclear how these probability vectors could be integrated into possible worlds.

Erk & Boleda 2014, Fu et al. 2014), though there is some debate on the extent to which particular methods allow for hypernymy inference (O. Levy et al. 2015). And work on mapping distributional representations to low-level visual features (Lazaridou, Bruni & Baroni 2014) is also relevant if we consider those visual features as cues towards visual properties such as colors.

Properties of words can also be induced more directly from distributional data, for example from the sentence *He saw the dangerous alligator* one can learn that alligators are sometimes dangerous (Almuhareb & Poesio 2004, 2005, Devereux et al. 2009, Kremer & Baroni 2010, Baroni, B. Murphy, et al. 2010, Baroni & Lenci 2010). It is difficult, with this technique, to learn the kinds of properties that humans give in feature norm experiments. Devereux et al. (2009) conclude that high-accuracy extraction of properties is unrealistic at this point in time. But it may be that extraction of exactly the same properties that humans tend to list is not what these models are good at. Baroni, B. Murphy, et al. (2010) observe the "tendency of [their model], when compared with the human judgments in the norms, to prefer actional and situational properties (riding, parking, colliding, being on the road) over parts (such as wheels and engines)" (p. 233). Thill, Pado & Ziemke (2014) suggest that the distributional features in models like the one of Baroni & Lenci (2010), which list closely co-occurring words, can reveal the "human experience of concepts." So it is possible that (some) distributional features themselves should also be considered as properties. This is an issue that I will not pursue further in the current paper, but that should be taken up again later.

5 A probabilistic information state

As discussed in the introduction, the two main aims of this paper are first, to argue that distributional inference (by the right kind of model) is property inference, and second, to propose a probabilistic inference mechanism for integrating distributional evidence with formal semantics. Section 4 addressed the first aim. Section 6 will address the second aim. This section lays the groundwork for Section 6 by defining *probabilistic information states*. If an agent is completely sure that crocodiles are animals, we will express this by having the probabilistic information state assign a probability of 1 to the statement *crocodiles are animals*. If the agent is unsure whether it is true that *alligators are animals*, their probabilistic information state will assign a probability of less than 1 to this statement (but greater than 0, because that would mean that the agent is sure that alligators are not animals).

5.1 Probabilistic semantics

I will give a probabilistic account of an agent's information state in terms of probabilistic logic (Nilsson 1986), which has a probability distribution over worlds to indicate uncertainty about the nature of the actual world. Probabilistic logic is easy to formulate in the case of finitely many worlds, but even though Clarke & Keller (2015) formulate it for the infinite case, it is not clear at this point how updates to the probability distribution extend to the infinite case. I argue below in Section 5.3 why I use probabilistic logic anyway. For now I use a definition that only allows for finitely many worlds.

As a basis for probabilistic logic, we use standard first-order logic languages L with the following syntax. Let IC be a set of individual constants, and IV a set of individual variables. Then the set of terms is defined as $IC \cup IV$. Let PS_n be a collection of n-place predicate symbols for $n \geq 0$. The set of formulas is defined inductively as follows. If R is an n-ary predicate symbols and $t_1, \ldots, t_n$ are terms, then $R(t_1, \ldots, t_n)$ is a formula. If ϕ_1, ϕ_2 are formulas, then so are $\neg\phi_1$ and $\phi_1 \wedge \phi_2$ and $\phi_1 \vee \phi_2$. If x is a variable and ϕ is a formula, then so are $\forall x\phi$ and $\exists x\phi$. A sentence of L is a formula without free variables. A *probabilistic model* for L is a structure $\mathcal{M} = \langle \mathcal{U}, \mathcal{W}, \mathcal{V}, \mathcal{P} \rangle$ where $\mathcal{U}$ is a nonempty universe, $\mathcal{W}$ is a nonempty and finite set of worlds,[6] and $\mathcal{V}$ is a valuation function that assigns values as follows. It assigns individuals from $\mathcal{U}$ to individual constants from IC. To an n-place predicate symbol $u \in PS_n$ it assigns a function $\mathcal{W} \to \mathcal{U}^n$. $\mathcal{V}(u)(w)$ is the extension of u in the world w. $\mathcal{P}$ is a probability distribution over $\mathcal{W}$. Then the probability of a sentence φ of L is the summed probability of all worlds that make it true:

$$P(\varphi) = \sum_{w \in \mathcal{W}} \{\mathcal{P}(w) \mid [\![\varphi]\!]^w = T\} \tag{7}$$

5.2 A probabilistic information state

We use probabilistic logic for describing the information state of an agent. Information states have been used in update semantics (Veltman 1996), where an information state is a set of worlds that, as far as the agent knows, could

6 Systems that work with probabilistic logic in practice, such as Markov Logic Networks (Richardson & Domingos 2006), usually assume a finite domain $\mathcal{U}$ that is in one-to-one correspondence with a finite set IC of constants, and a finite set of predicate symbols. But minimally, what is needed is that the set of worlds $\mathcal{W}$ be finite.

be the actual world. In a probabilistic information state, the agent considers some worlds more likely than others to be the actual world (van Benthem, Gerbrandy & Kooi 2009, van Eijck & Lappin 2012, Zeevat 2013).

We define a probabilistic information state over language L to be a probabilistic model $\mathcal{M}$ for L. While update semantics focuses on the way that a single sentence updates the information state, we are interested in describing how the probability distribution over worlds is affected by distributional information when that distributional information contains clues about the meaning of a word u. After all, the probability that an agent ascribes to a world depends, among other things, on the agent's belief about what words mean.

For example, assume that the distributional similarity of *crocodile* and *alligator* is 0.93. We will call this a piece of distributional evidence E_{dist}. 0.93 is a high similarity rating, so we would be likely to see this evidence if the actual world is a world w_1 in which entities that are alligators and entities that are crocodiles have many properties in common. We would not be so likely to see this evidence if the actual world was a world w_2 where alligators and crocodiles do not have many properties in common. In our case, the agent should assign a higher probability to w_1 and assign a lower probability to w_2 on the basis of E_{dist}. This can be described through Bayesian belief update, which in this case would use the probability of the distributional evidence in a world to compute the inverse: the probability of a world given the distributional evidence. Bayesian update transforms a *prior* belief, in our case a prior probability distribution P_0 over worlds, into a *posterior* belief, in our case a probability distribution P_1 that describes the probability of each world w after the distributional evidence has been seen. It does so by weighting the prior probability of w against the probability of the evidence in world w, and normalizing by the probability of the evidence.

$$P_1(w) = P(w|E_{\text{dist}}) = \frac{P(E_{\text{dist}}|w)P_0(w)}{P(E_{\text{dist}})} \tag{8}$$

The probability of the evidence $P(E_{\text{dist}})$ in the denominator of (8) is computed as the sum of the probability of E_{dist} under all worlds w' weighted by the probabilities of those worlds:

$$P(E_{\text{dist}}) = \sum_{w'} P(E_{\text{dist}}|w')P_0(w') \tag{9}$$

Equation (8) contains three different probabilities: the prior P_0, which we take to be given, the posterior P_1, which we compute, and the probability

$P(E_{\text{dist}}|w)$ of the distributional evidence in a world w; Section 6 will describe how an agent can compute this third probability. There, we will formulate distributional evidence in the form of an inequality: "The distributional similarity of words u_1 and u_2 is as high as s or higher," for reasons that I explain in Section 6.

In update semantics, the information state is updated through every sentence that the agent hears. Equation (8) suggests an update on the probabilistic information structure, from $\langle \mathcal{U}, \mathcal{W}, \mathcal{V}, P_0 \rangle$ to $\langle \mathcal{U}, \mathcal{W}, \mathcal{V}, P_1 \rangle$. But in the case of distributional evidence, we do not model a dynamic system, that is we do not imagine that there is a particular point at which an agent decides to learn from all the accumulated distributional evidence about alligators and crocodiles. Rather we want to treat the distributional contribution to word meaning as static. (This is of course a simplification, as the distributional data changes with every sentence that an agent hears.) So the formulation in (8) is a useful modeling device, but is somewhat misleading, as it suggests a change in the information state. There is a second, equivalent formulation that is more appropriate. Suppose that, instead of having explicit access to the probability of each world in the distribution P_0, we approximate P_0 by repeatedly sampling from the collection of worlds in such a way that the probability of drawing w corresponds to $P_0(w)$. (This can be done, and I show an example in Section 7.) Then it is possible to sample from the posterior distribution P_1 of Equation (8) instead of the prior P_0 through a condition on the sampling. The idea is that we sample a world w according to P_0, but then only keep it if a condition is met. We use w to generate a piece of "pseudo-distributional evidence" s': What should the distributional similarity rating look like if the actual world was w? If this s' is reasonably similar to the actual similarity that was observed in E_{dist}, we retain w as sampled for the posterior P_1, otherwise we reject it. I discuss this in more detail in Sections 6 and 7.

5.3 Alternatives to probabilistic logic

This paper needed a probabilistic mechanism for representing uncertain information and inference from uncertain information in an agent, and I chose to use probabilistic logic. As I mentioned above, probabilistic logic is straightforward to define in the case of finitely many worlds, but it is difficult to extend to the case of infinitely many worlds. There are some potential alternatives to probabilistic logic, but I am now going to argue that none

of them is currently viable for my purposes. So like Zeevat (2013), I choose to use probabilistic logic in spite of its problems because it allows us to explore the connection of probabilities and logic in a framework that is close to standard model-theoretic semantics.

Fuzzy logic (Zadeh 1965) is a many-valued logic in which formulas are assigned values between 0 and 1 that indicate their degrees of truth, and it can represent uncertainty in that way. Fuzzy logic is a truth-functional approach, as the truth value of a complex formula is a function of the truth values of its components. For example, the truth value of a formula $F \wedge G$ is the minimum of the values of F and G. The problem with truth-functional approaches is that they miss penumbral connections (Fine 1975, van Deemter 2013), like the fact that an object cannot be all pink and all red at the same time. If an object b is on the borderline of pink and red, the weights for "b is pink" and "b is red" could both be 0.5. In that case the value for "b is pink and b is red" will be 0.5 as well, while it should be zero (false). Because it produces counter-intuitive truth degrees of this kind, I would rather use more principled alternatives to fuzzy logic.

Cooper et al. (2014) focus on the interaction between perception and semantic interpretation. They use a type-theoretic setting, in which situations are judged as being of particular types, which can be arbitrary propositions (among other things). Probabilities are associated with these judgments. The probabilities are assumed to come from a classifier that takes as input either some feature-based perceptual representation of a situation or a collection of previous types assigned to the same situation, and classifies it as being of a particular type with a certain probability. So penumbral connections will be taken into consideration only to the extent that the classifier has learned to respect them. That is, if the classifier never sees a situation in which something is completely pink and completely red at the same time, the probability of "b is pink and b is red" will be zero – but this information can only be learned by the classifier, and the classifier is a black box. No declarative knowledge can be injected into it. So I do not use the framework of Cooper et al. because it would constrain me to knowledge collected by direct perception, while I want to study the interaction of declarative knowledge with information coming from distributional evidence.

Goodman, Tenenbaum & Gerstenberg (to appear) and Goodman & Lassiter (2014) propose that in understanding an utterance the hearer builds a mental situation. So their approach is situation-based like the one of Cooper et al., but their situations are imagined, while Cooper et al. use perception of real

situations. The hearer generates a mental situation based on probabilistic knowledge (where Goodman et al. do not make a distinction between general world knowledge and linguistic knowledge) encoded in probabilistic *generative* statements like: "To generate a person, draw their gender by flipping a fair coin, then draw their height from the normal distribution of heights for that gender." I do not use their approach because it focuses on generating mental situations based on individual utterances, and the current paper is not about updating the information state based on individual utterances. But generative models are widely used in machine learning (Bishop 2006), and the idea of using a generative approach to assemble a situation or world probabilistically (which is introduced in more detail below) is general enough that it can be used to implement the model introduced in this paper. It is just that in our case what is generated is not a situation but a world. Goodman, Mansighka, et al. (2008) have developed a programming language for specifying (small) probabilistic generative models, which is general-purpose and not restricted to implementing the specific model of Goodman et al. I use it for a proof of concept experiment in Section 7.

6 Probabilistically inferring properties from distributional evidence

This section introduces a mechanism by which distributional inference can influence the probabilistic information state. In the previous section I have described the effect of distributional evidence as a (Bayesian) update on a probability distribution over worlds, with a formula that is repeated here for convenience: The prior belief $P_0(w)$, or probability that the agent assigns to w being the actual world, is transformed into a posterior belief $P_1(w) = P(w|E_{\text{dist}})$ after seeing the distributional evidence E_{dist}, which is about a distributional similarity rating for a particular pair of words:

$$P_1(w) = P(w|E_{\text{dist}}) = \frac{P(E_{\text{dist}}|w)P_0(w)}{P(E_{\text{dist}})} \tag{8}$$

The Bayesian update in Equation (8) has three components: the prior probability $P_0(w)$ of the world w, the posterior probability $P_1(w)$ of w after seeing the evidence E_{dist}, and the probability of the distributional evidence in the world w, $P(E_{\text{dist}}|w)$. (The denominator, $P(E_{\text{dist}})$, can be factored into the same components, as shown in the previous section.) The prior is given. To obtain the posterior, we need to compute $P(E_{\text{dist}}|w)$. This section shows how that can be done.

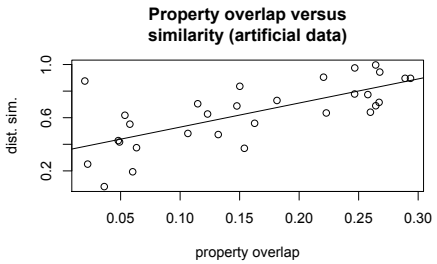
6.1 The story in a nutshell

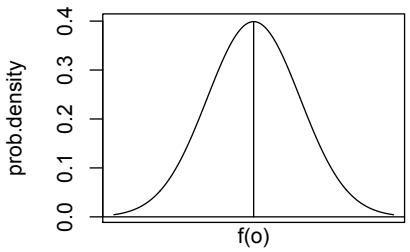
Before I go into the details of the model, I first give an overview of its pieces. Each distributional similarity value is a weight, and to use it as evidence, the agent first needs to interpret it: What kinds of distributional similarity values typically occur when property overlap is high? And for low property overlap? The top left panel of Table 4 gives an example. It shows pairs of words for which we assume the agent knows the property overlap ("ovl"), along with the distributional similarity ("sim") that the agent observes for them. (Section 6.4 below defines what I mean by "knowing the property overlap.") What the agent can observe in this case is that high property overlap tends to go with high distributional similarity, and vice versa. This observation can in the simplest case be described as a linear relation, as sketched in the top right panel of Table 4.

When the agent has inferred this linear relationship, they can take a property overlap o between the extensions of two words u_1 and u_2 and predict what the distributional similarity $f(o)$ of u_1 and u_2 should be, simply by reading off the appropriate y-value (similarity) for the given x-value (overlap). Here, $f(.)$ is a function that maps an overlap value o to its fitted value, the model's prediction for o. And the agent can go one step further: The data is somewhat noisy, in that the distributional similarity is usually not exactly equal to the value $f(o)$. But it will most likely be close to $f(o)$, and will be less likely to be much higher or much lower than that. This can be described as a probability distribution over possible similarity values given overlap o. It has its mean at $f(o)$, as sketched in the bottom left panel of Figure 4. Section 6.5 describes this probability distribution in more detail.

This probability distribution can then be used to estimate $P(E_{\text{dist}}|w)$, the probability that this section is about: Suppose the property overlap of the extensions of u_1 and u_2 in world w is o, the distributional similarity that the agent predicts from overlap o is $f(o)$, and the actually measured distributional similarity of u_1 and u_2 is s. What is the probability of this distributional evidence given w, which is to say, given that the property overlap is o? This can be read off the probability distribution in the bottom left panel of Figure 4, which is a probability distribution over the distributional similarity values we expect to see given that the overlap is o. But what we need is not the probability of seeing a similarity rating of exactly s. This probability does not say anything about whether s is high or low, as the probability distribution is symmetric, and the probability of seeing a similarity

word 1	word 2	ovl	sim
peacock	raven	0.29	0.70
mixer	toaster	0.19	0.72
crocodile	frog	0.17	0.86
bagpipe	banjo	0.10	0.72
scissors	typewriter	0.04	0.62
crocodile	lime	0.03	0.33
coconut	porcupine	0.03	0.42





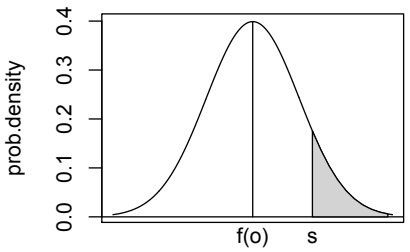


Figure 4 Estimating $P(E_{\text{dist}}|w)$: Property overlap (ovl) and distributional similarity (sim) observed for words with known properties (top left), and the regularity described through linear regression (top right). Bottom: distribution of expected similarity ratings centered on linear regression prediction $f(o)$, the fitted value for o (left), and probability of seeing a rating as high or higher than the observed s (right)

rating of $f(o) + d$, for any value d, is the same as seeing a similarity rating of $f(o) - d$. But the probability of seeing a similarity rating that is as high as s or higher – drawn as the shaded area in the bottom right panel of Figure 4 – does give us an indication on whether s is high or low: The lower s is, the bigger is the shaded area. So I will formulate the distributional evidence E_{dist} as "the distributional similarity of u_1 and u_2 is as high as s or higher".

This method for determining $P(E_{\text{dist}}|w)$, which is described in detail in Section 6.6, can be characterized as *hypothesis testing*. The hypothesis that we are testing is that w is the actual world, and we are testing it against the evidence E_{dist}. The hypothesis that w is the actual world is associated with a probability distribution, namely the one in the bottom left panel of Figure 4: If w were the actual world, then the most likely distributional similarity we would expect to see for u_1 and u_2 is $f(o)$, with other values having lower and lower probability the further they are from $f(o)$. Now we determine $P(E_{\text{dist}}|w)$, the probability of seeing the similarity of u_1 and u_2 be as high as s or higher, under the assumption that the hypothesis (that w is the actual world) is correct. If $P(E_{\text{dist}}|w)$ turns out to be low, then we have reason to reject our initial hypothesis, or in our case, we have reason to assign a lower probability to the hypothesis that w is the actual world.

With $P(E_{\text{dist}}|w)$ in hand, we can compute the posterior $P(w|E_{\text{dist}})$, the probability that the agent ascribes to w being the actual world after seeing the distributional evidence E_{dist}, according to Equation (8). This is a probabilistic information state, one that is informed by distributional evidence. At the beginning of the paper I raised the question of whether an agent can infer that "φ_1: all alligators are dangerous," "φ_2: all alligators are edible," or "φ_3: all alligators are animals" based on distributional evidence. What I meant, more precisely, is this: In the (posterior) probabilistic information state that is informed by distributional evidence, will the agent ascribe a higher probability to φ_1, φ_2, or φ_3 than in the prior probabilistic information state that is unaware of the distributional evidence?

The approach that I just sketched assumes a two-way interaction between the knowledge that an agent has, as encoded in their probabilistic information state, and the distributional evidence. The agent uses the knowledge that is in the probabilistic information state to interpret distributional data (upper two panels of Figure 4). Conversely, the distributional evidence can then influence the probabilistic information state (lower two panels of Figure 4).

6.2 Properties

Distributional models produce similarity ratings for pairs of target words. In our case, as mentioned above, we focus on nouns and we assume for now that there is no polysemy. We assume that each target word u corresponds to an n-place predicate u' for some n. This predicate u' is interpreted in the agent's probabilistic information state through the interpretation function $\mathcal{V}$, where $\mathcal{V}(u')$ is a function from worlds to extensions (sets of n-tuples of entities). We will also take properties to be functions from worlds to extensions. In this paper, we adopt a broad definition of a property: The interpretation of any n-place predicate can be a property. A property does not have to be the interpretation of an adjective; it can be the interpretation of a noun or a verb. Importantly, hypernyms are also properties, as in "all alligators are animals."

For a given world w, we say that the extension of u' in w *possesses the property* q if it is included in the extension of q in w: $\mathcal{V}(u')(w) \subseteq q(w)$. Note that on this definition, the extension of q has to have the same arity as the extension of u', or the extension of u' cannot possess the property q.

6.3 Uncertainty about extensions and about properties

We need to distinguish two types of uncertainty about word meaning: uncertainty about the extension, and uncertainty about properties. An agent is *uncertain about the extension* of the word u in the following case:

- The agent knows that the word exists, knows that it is a noun, can successfully syntactically analyze sentences that include the word u, and is able to represent it by an n-place predicate symbol u' for an appropriate n.
- But the extension of u' varies across worlds to which the agent ascribes nonzero probability. Formally, there are worlds w_1, w_2 that are both ascribed nonzero probability by the agent such that the extension of u' is not the same in w_1 and w_2:

  $\exists w_1, w_2 \in \mathcal{W}(\mathcal{P}(w_1) > 0 \wedge \mathcal{P}(w_2) > 0 \wedge \mathcal{V}(u')(w_1) \neq \mathcal{V}(u')(w_2))$.

It is important that in this definition of uncertainty we only need to take into account worlds w that have nonzero probability in the agent's information state. The worlds to which the agent ascribes a probability of

zero do not affect the agent's certainty about what the word u means. We will say that an agent *takes the world w into consideration* if w has a nonzero probability in the agent's information state.

The uncertainty that we are centrally concerned about in this paper is not uncertainty about extensions but *uncertainty about properties*. For example the agent may be unsure whether all alligators are animals or not. We will say that the agent is uncertain whether the word u (represented by a predicate u') has property $\mathcal{V}(v')$ if in some worlds that the agent takes into consideration, the extension of u' is included in the extension of v', and in some other worlds that is not the case: $\exists w_1, w_2 \in \mathcal{W}(\mathcal{P}(w_1) > 0 \wedge \mathcal{P}(w_2) > 0 \wedge \mathcal{V}(u')(w_1) \subseteq \mathcal{V}(v')(w_1) \wedge \mathcal{V}(u')(w_2) \nsubseteq \mathcal{V}(v')(w_2))$.

If an agent is uncertain about whether all alligators are animals, this could be either because the agent is uncertain about the extension of *alligator* or because the agent is uncertain about the extension of *animal* – or both. In either case, doing inference based on the distributional similarity of *alligator* and *crocodile* (assuming the agent is certain that all crocodiles are animals) should reduce the agent's uncertainty about alligators being animals. Below I ignore the question of whether the agent is certain or uncertain about the extensions of words, and focus on the question of uncertainty about properties. (Though reducing an agent's uncertainty about the properties of alligators can also reduce the agent's uncertainty about the extension, in that it restricts the extension of *alligator* to being a subset of the extension of *animal*.)

6.4 Property overlap

We measure property overlap through the Jaccard Coefficient (Jaccard 1901), which is defined as the intersection size of two sets relative to the size of their union. The Jaccard similarity of two sets s, t is

$$\text{Jaccard}(s,t) = \frac{|s \cap t|}{|s \cup t|} \tag{10}$$

In our case, we want the Jaccard Coefficient to measure the degree to which the sets of properties of two extensions $\mathcal{V}(u')(w)$ and $\mathcal{V}(v')(w)$ overlap. Jaccard is defined only for finitely many properties. So for simplicity we assume some finite set of relevant properties $Prop \subseteq \bigcup_n \mathcal{W} \to \mathcal{U}^n$. (Here and below we assume that all probabilistic information states that the agent can assume have the same universe, the same set of worlds and the same

valuation function, and only differ in the probability distribution over worlds. For that reason, the model structure $\mathcal{M}$ is not used as an index in definitions. When a definition depends on the probability distribution over worlds, $\mathcal{P}$ is used as an index.) We write $PropOf(\mathcal{V}(u'), w)$ for the properties in *Prop* that $\mathcal{V}(u')$ possesses in world w. Now we can describe the property overlap between the extensions of two predicate symbols u', v' in a given world w as the Jaccard Coefficient over their respective sets of properties in w.

(11)
Property overlap in world w:

$$\mathrm{Ovl}(\mathcal{V}(u'), \mathcal{V}(v'), w) = \frac{|\, PropOf(\mathcal{V}(u'),w) \cap PropOf(\mathcal{V}(v'),w)\,|}{|\, PropOf(\mathcal{V}(u'),w) \cup PropOf(\mathcal{V}(v'),w)\,|}$$

The properties that the extension of u' possesses can vary across worlds. Alligators may be edible in some worlds, and not have that property in other worlds. We will say that the agent is *certain* that $\mathcal{V}(u')$ has property q, or that q is a *reliable property* of $\mathcal{V}(u')$, if the extension of u' is included in the value of q in any world that the agent takes into consideration.

Reliable properties:

$$RPropOf_{\mathcal{P}}(\mathcal{V}(u')) =$$
$$\{q \in Prop \mid \forall w \in \mathcal{W}\big(\mathcal{P}(w) > 0 \Rightarrow \mathcal{V}(u')(w) \subseteq q(w)\big)\}$$

We can now define *reliable property overlap* as property overlap involving only reliable properties. One reason why we need reliability is that, as the two topmost panels of Table 4 show, the agent is supposed to estimate a relation between degree of property overlap on the one hand and distributional similarity on the other. If property overlap varies widely across worlds, which property overlap should the agent use for this estimation? But reliable properties alone are not enough to solve the problem. If the agent has no idea what alligators are, except that they are alligators (self-identity is always a property of which the agent is certain), then the property overlap of *alligator* and *crocodile* will be unreliable even when restricted to reliable properties. So we assume that the agent uses some threshold θ, for a value of θ much larger than 1, and will consider reliable property overlap to be defined only if

alligator	
has_teeth	has_a_tail
is_green	is_scary
an_animal	swims
is_long	has_a_mouth
a_reptile	lives_in_Florida
lives_in_swamps	has_jaws
has_scales	eats_people
lives_in_water	is_dangerous

Table 3 Example from the McRae et al. 2005 feature norms: Properties of *alligator* collected from human subjects.

both $\mathcal{V}(u')$ and $\mathcal{V}(v')$ have at least θ reliable properties.

Reliable property overlap:

$$\text{ROvl}_{\mathcal{P}}(\mathcal{V}(u'), \mathcal{V}(v')) = \begin{cases} \frac{|\,RPropOf_{\mathcal{P}}(\mathcal{V}(u')) \cap RPropOf_{\mathcal{P}}(\mathcal{V}(v'))\,|}{|\,RPropOf_{\mathcal{P}}(\mathcal{V}(u')) \cup RPropOf_{\mathcal{P}}(\mathcal{V}(v'))\,|} & \text{if } |RPropOf_{\mathcal{P}}(\mathcal{V}(u'))| \geq \theta \\ & \text{and } |RPropOf_{\mathcal{P}}(\mathcal{V}(v'))| \geq \theta \\ \text{undefined} & \text{else} \end{cases} \tag{12}$$

For example, suppose that for a particular agent, the reliable properties of crocodiles are that they are crocodiles, animals, aquatic, and green. And suppose that for the same agent, the reliable properties of seaweed are that it is seaweed, aquatic, and green. If $\theta = 3$, then we have a reliable property overlap of *crocodile* and *seaweed* of $\frac{2}{5}$. However, if $\theta = 4$, then that means that the agent knows too little about *seaweed*, and the reliable property overlap of *crocodile* and *seaweed* is undefined.

6.5 Predicting similarity from property overlap

At this point we have all the definitions we need to flesh out the story in Figure 4. The first step, illustrated in the top left panel, is for the agent to compare the *reliable property overlap* between $\mathcal{V}(u')$ and $\mathcal{V}(v')$ to the distributional similarity of the corresponding words. The property overlap values (in the "ovl" column) in that top left panel were computed from the McRae et al. 2005 feature norm data. As briefly mentioned in Section 3, these are human-generated definitional features. As an example, Table 3 shows

features collected for *alligator* in the McRae et al. dataset.[7] All and only the features listed in the feature norm dataset were taken to be reliable properties (this assumption is in fact false, but we just have to consider this as noise in the data), and property overlap was computed as the Jaccard coefficient.[8] The distributional similarities (in the "sim" column) were computed using the distributional model defined in Section 2.4, a model with narrow context window and cosine similarity. This combination of the McRae et al. feature norm data and the distributional model from Section 2.4 will be our running example in this and the following section.

Observing the reliable property overlap and distributional ratings, the agent will note that higher property overlap tends to go with higher similarity. This observation can be made precise through linear regression, as illustrated in the top right panel of Figure 4. Linear regression infers the line that best describes the linear relationship between two variables, X and Y. Any line can be described through two parameters: The intercept β_0 indicates the y-value at which the line crosses the y-axis, and the slope β_1 indicates the amount by which the line rises or falls as the y-value increases by one. Given β_0 and β_1, the y-value for a given x can then be predicted as the intercept plus the slope multiplied by x:

$$Y = \beta_0 + \beta_1 \cdot X \tag{13}$$

The values β_0 and β_1 can be learned from data consisting of pairs of observed x-values and observed y-values, as in the upper left panel of Figure 4. As the points will never form a perfect line, the model makes the assumption that the actually observed y-values Y_{obs} (as opposed to the values Y, which are predicted from X rather than observed) depend on the x-values through a linear relation plus some error ε that makes the y-values deviate from the perfect line:

$$Y_{obs} = \beta_0 + \beta_1 X + \varepsilon \tag{14}$$

Then the intercept and slope are computed as the values that minimize the square of the error ε. In our case, we want to predict similarity from property overlap: If two extensions have a property overlap of o, what is the most likely distributional similarity value $f(o)$ between the corresponding words

7 The researchers compiling feature norm data normalize the feature norms after collection, such that for example "green" and "is green" would be counted as the same feature.

8 For any concept c, I added the property of "being a c."

going to be? So equation (13) becomes

$$f(o) = \beta_0 + \beta_1 o \tag{15}$$

I trained a linear regression models on pairs of reliable property overlap and distributional similarity, $\langle \text{ROvl}_P(\mathcal{V}(u'), \mathcal{V}(v')), sim(u, v)\rangle$, again using the feature norms of McRae et al. 2005 and the distributional model from Section 2.4.[9] All word pairs in which one of the two words was *alligator* were omitted from the training data, as we are going to use *alligator* as a word about whose properties we are uncertain in the experiments in the next section. The linear regression model has an intercept of $\beta_0 = 0.35$ and slope of $\beta_1 = 1.80$ and explains about 20% of the variance in the data (adjusted R-squared value of 0.205). The amount of variance explained is not immensely high, but that is to be expected given that the feature norms do not list all the properties that apply to a given concept – in fact they omit many.

With equation (15), an agent can take the property overlap o between the extensions of u' and v' in some world w and predict what the most likely distributional similarity $f(o)$ of u and v should be. But as sketched in the lower left panel of Figure 4, the agent can do even more, in particular the agent can infer a probability distribution over possible similarity values given that the property overlap is o. The most likely similarity value is $f(o)$, but the value could also be somewhat higher or somewhat lower. And in fact, the linear regression model, in particular the error term ε in (14), can be used to estimate how widely the similarity values will tend to vary around $f(o)$: If the observed similarity values are always very close to the predicted value, then the distribution of similarity values should be rather peaked; if the observed similarity values are often far higher or far lower than the predicted value, the distribution of similarity values should be broader. Linear regression modeling makes the assumption that the error is normally distributed around the the predicted value. Based on this, we can describe the observed value Y_{obs} as drawn from a normal distribution with the predicted value as its mean:

$$Y_{obs} \sim N(\beta_0 + \beta_1 X, \text{sdev}^2) \tag{16}$$

The standard deviation in equation (16) is the residual standard error of the linear regression model. For our running example, this value is sdev =

9 The linear regression model was computed using the `lm` function in the R statistics package, R version 3.1.2 (R Core Team 2014).

0.1481. So the agent's predictions about what the distributional similarity will be, given two extensions with property overlap o, can be described as a normal distribution with mean $f(o)$. We write this probability density as $g_o = N(f(o), \text{sdev}^2)$.

6.6 The probability of a piece of distributional evidence

A piece of distributional evidence E_{dist} is about the observed distributional similarity s of two words u and v. As discussed above, we do not formulate E_{dist} as "seeing a similarity value that is exactly s" because the normal distribution is symmetric, and the probability of seeing a high value of $f(o) + d$, for some d, is the same as the probability of seeing a low value of $f(o) - d$. Instead, we formulate E_{dist} as "seeing a similarity value as high as s or higher." The higher the observed similarity s, the lower the probability of a piece of distributional evidence, as can be seen in the lower right panel of Figure 4. A piece of distributional evidence involves two words u and v and their similarity $sim(u, v) = s$, so it has the form

$$E_{\text{dist}} = \langle u, v, \geq sim(u, v) \rangle \tag{17}$$

The probability of E_{dist} can be expressed in terms of the *cumulative distribution function* G_o of g_o. $G_o(s)$ is the probability of seeing a value as low as s or lower with the probability density g_o: $G_o(s) = \int_{-\infty}^{s} g_o(x)dx$. As $G_o(s)$ is the probability of "as low as s or lower", what we will need is $1 - G_o$. The probability of the distributional evidence E_{dist} given w is the cumulative probability of seeing a similarity value as high as s or higher if the predicted similarity value for the observed property overlap o is $f(o)$:

(18) *Probability of observed distributional evidence* $E_{dist} = \langle u, v, \geq s \rangle$ *in world* w *if* $Ovl(\mathcal{V}(u'), \mathcal{V}(v'), w) = o$*:*

$$P(E_{\text{dist}} \mid w) = 1 - G_o(s)$$

This probability $P(E_{\text{dist}} \mid w)$ can be used in the Bayesian update equation (8) to compute the posterior probability $P_1(w) = P(w|E_{\text{dist}})$ from the prior probability $P_0(w)$, where $P_1(w)$ is the probability of the world w after the distributional evidence has been taken into account.

We use the Bayesian update (Equation 8) because it is a well-known and straightforward way to describe the influence of evidence, in this case distributional evidence, on an agent's beliefs. But there is a second equivalent

option for describing the influence of distributional evidence on the probabilistic information state, namely as a sampling condition. I go to the trouble of using this second formalization for several reasons. First, Bayes' rule suggests that there is a single moment of belief update, but that is not an assumption that we make, as discussed in Section 5. Second, I will implement distributional inference through sampling conditions in the experiments in Section 7. Third, sampling conditions allow us to describe the influence of distributional evidence in a declarative fashion that can be described as a *soft meaning postulate*. Standard meaning postulates can be viewed as placing a constraint on the set of worlds that could possibly be the actual world, where that constraint is hard: Any world that does not conform to the meaning postulate cannot be the actual world. Soft meaning postulates instead place a soft constraint on the probabilistic information state, where worlds that do not meet the constraint will have lower probability.

The idea is that if P_0 can be approximated by sampling, then P_1 can be approximated by imposing an additional condition on the sampling. Approximating P_0 by sampling means repeatedly sampling worlds (with replacement) from the set $\mathcal{W}$ in such a way that the probability of drawing a world w approximates $P_0(w)$. In Section 7 we approximate P_0 by a sampling procedure that probabilistically generates worlds, where "generating a world" means assigning properties to entities in a fixed-size universe through a random process. To approximate P_1, we add a condition that is specific to a piece of distributional evidence $E_{\text{dist}} = \langle u, v, \geq s \rangle$. Once a world w has been sampled according to P_0, we generate a piece of "pseudo-distributional evidence" s' from w: We determine the property overlap $\text{Ovl}(\mathcal{V}(u'), \mathcal{V}(v'), w) = o$ of the extensions of u' and v' in w, and draw a similarity value s' from the distribution g_o. If s' is greater or equal to the observed similarity s from E_{dist}, we keep w in our sample, otherwise we reject it:

(19)
Soft meaning postulate of $E_{\text{dist}} = \langle u, v, \leq s \rangle$ *in world* w
if $Ovl(\mathcal{V}(u'), \mathcal{V}(v'), w) = o$*:*

$s' \geq s \quad \text{for } s' \sim g_o$

6.7 Discussion

Talking about alligators. We return once more to the question of how it is possible that an agent would successfully use a word without being aware of its extension. At this point, we can answer this question based on the two

different definitions of uncertainty in Section 6.3. An agent may be unable to point out an alligator, in the sense that the extensions of *alligator′* vary widely across the worlds that the agent takes into consideration (which I defined above as worlds that have non-zero probability in the agent's information state), and the agent can still be absolutely certain that all alligators are animals, in the sense that the extension of *alligator′* is a subset of the extension of *animal′* in all worlds that the agent takes into consideration. In that case, the probability of "all alligators are animals" in the agent's probabilistic information state is 1, and the agent can confidently make statements based on the fact that alligators are animals. The same holds for banyan trees and elms (two examples mentioned in the introduction).

Suppose I am unable to point out an alligator in a zoo, but I am certain that alligators are animals, and I make statements about alligators based on my information state. This creates distributional data that other speakers might want to use to learn about word meaning. And in fact if these other speakers now want to learn about properties of alligators, they can confidently use the distributional data that I produced, even though I have no idea about the extension of *alligator′*: If I am certain that alligators are animals, this will be reflected in the verbs and modifiers that I apply to the word *alligator*, which in turn will create good data for other speakers to learn from.

Distributional information and formal semantics, loosely coupled through inference. In this section I have proposed a mechanism for integrating distributional property inference with formal semantics. This mechanism couples two different formalisms loosely through probabilistic inference, so it is not a single unified model. And in fact I would say that the eventual aim of an integration of formal semantics with distributional information would not necessarily have to be a unified model. Meaning is a complex phenomenon. It involves a notion of truth and grounding, and formal semantics captures that well. In this paper I have argued that meaning can also be learned from observed use in a community of speakers, which is here represented distributionally. And this observed use is not centrally about truth: Useful lexical information can be learned from some non-factual statements. For example, from "Alligators love to eat bananas" an agent can infer that alligators are likely to be animate (as they can be the subject of *eat*). So as truth is not central to distributional data, it makes sense to represent it differently.

Cognitive plausibility of the model? In this section I have proposed an approach for how distributional data could be used to influence an agent's information state. Now I briefly turn to the question of whether it would be plausible for a human agent to use distributional data (which is not the same as to prove that human agents use distributional data; I am not sure that the latter would be possible.) The strongest and most well-known argument that speakers do use distributional information was made by Landauer & Dumais (1997), who proposed distributional models as "a solution to Plato's problem". This is the question of how humans can possibly acquire the large vocabulary that they command, and Landauer and Dumais' answer is that humans use contextual clues to learn the meaning of words that they observe in linguistic input. Some experimental support for the hypothesis of distributional word learning comes from McDonald & Ramscar (2001). They use context to influence the perceived meaning of either marginally familiar words or made-up words, with a highly significant effect on subsequent similarity ratings.

Another argument is that many phenomena in semantics seem to involve gradedness and that distributional models with their graded notion of similarity provide a mechanism to explain how humans are able to make such graded judgments. This argument can be found in van Eijck & Lappin 2012, though they reject distributional models in the end. Synonymy is an example of a phenomenon that seems to involve gradedness: Instead of absolute synonymy, we find near-synonymy of words that are often substitutable but still differ in nuances of meaning (Edmonds & Hirst 2002). Polysemy also seems to come in degrees, with different uses of a word differing in their perceived similarity (Brown 2008, Erk, McCarthy & Gaylord 2013). However, distributional similarity of observed textual contexts is not the only possible source for graded semantic judgments. Other possibilities include similarity of semantic feature representations, or a probabilistic match between an entity or situation on the one hand and a label on the other hand (Cooper et al. 2014). Still, it is possible that distributional similarity would be one of the mechanisms involved.

A third argument, proposed by Baroni, Bernardi & Zamparelli (2014), is simply that distributional evidence is available in ample amounts and it is demonstrably useful, so it stands to reason that humans would make use of it.

Taken together, these arguments from the literature make the case that it is reasonable to assume that speakers make use of distributional evidence.

They do not prove that speakers necessarily use this kind of information, but as I said above, this stronger argument is not one I am even attempting to make.

The converse question is whether there are any reasons why humans cannot make use of any distributional clues. There does not seem to be any work that argues this point, but there are many papers, for instance van Eijck & Lappin 2012, that argue a weaker point, namely that not all of human language processing can be distributional. Lenci (2008) lists three main reasons: that distributional models lack compositionality, do not support inference, and cannot provide reference. Although these three problems remain difficult for distributional models, there has been considerable work on all of them since Lenci's paper appeared. Concerning compositionality, the compositional construction of distributional phrase representations is currently an active area of research (Landauer & Dumais 1997, Mitchell & Lapata 2010, Baroni & Zamparelli 2010, Coecke, Sadrzadeh & Clark 2011, Grefenstette & Sadrzadeh 2011, Socher et al. 2012, Baroni, Bernardi & Zamparelli 2014). We have discussed the topic of distributional inference in Section 3. And concerning reference, there has been work on vector space representations that integrate textual, visual, and conceptual information (Feng & Lapata 2010, Andrews, Vigliocco & Vinson 2009, Bruni et al. 2012) and on mapping between textual and visual spaces (Lazaridou, Bruni & Baroni 2014).

Summing up these arguments that aim to show that not all of human language processing can be distributional, what they mostly show is that distributional models are a moving target. The field is currently making fast progress, such that it is hard to say which constraints on the expressive power of distributional models will turn out to be fundamental and which will be overcome by future models. Still, the picture that emerges from the current limitations of distributional models is that distributional evidence is conceivable as part of a human language processing apparatus, but not all of it. This is also the view that I take in this paper.

7 Learning about alligators: three experiments

This section reports on three proof of concept experiments that showcase the mechanism from the previous section. The experiments are kept simple in order to better illustrate the properties of the model: In the first two experiments we study how the model learns from one single word, at either high or medium distributional similarity to the unknown word. In the third

experiment we test how the model learns from two different words at the same time.

We focus on the three example sentences from the introduction, repeated here for convenience:

φ_1: All alligators are dangerous.

φ_2: All alligators are edible.

φ_3: All alligators are animals.

We assume an agent who is uncertain about the properties of alligators, but certain about the properties of crocodiles and trouts. Among other things, the agent is certain that crocodiles are dangerous, trouts are edible, and both crocodiles and trouts are animals, in the sense of Section 6.3: In all worlds that the agent takes into consideration (to which they assign nonzero probability), the extension of *crocodile*′ is a subset of the extension of *dangerous*′, and analogously for the other properties.

In the first experiment, we use sentence φ_1 to test how distributional similarity of *alligator* and *crocodile* influences the degree of the agent's certainty that alligators are dangerous. The second experiment tests the influence of the distributional similarity of *alligator* and *trout* on the probability that the agent ascribes to φ_2. The difference between the first two experiments is that the distributional similarity of *alligator* to *crocodile* is very high while the similarity to *trout* is only moderate. With sentence φ_3 we test whether evidence (about alligators potentially being animals) can accumulate if we have distributional similarity ratings of *alligator* to both *crocodile* and *trout*. But let me point out again that the evidence in all three cases is probabilistic and noisy. We will not be able to conclude with certainty that all alligators are dangerous, or even that all alligators are animals. And that is as it should be: We do not want to draw inferences with complete certainty based on evidence that does not provide complete certainty. What I aim to show instead is that the distributional evidence significantly affects the agent's probabilistic information state, in particular that evidence about distributional similarity can bring the agent to consider it more likely that alligators have properties that crocodiles and trouts also possess. That is the measuring stick by which we will judge whether we have been able to show that distributional evidence is useful.

Throughout this section, we use sampling to approximate the prior probability distribution P_0 over possible worlds. And we use sampling constraints

in the form of the soft meaning postulates in Equation (19) to sample from the posterior P_1 instead of P_0, where P_1 is the probability distribution over possible worlds that takes the distributional evidence into account.

To obtain the numbers for our experiments, we again use the McRae et al. 2005 feature norms as property lists, we use the distributional model from Section 2.4, and we use the linear regression model from the previous section that predicted distributional similarity from property overlap.

Experiment 1: Are alligators dangerous? For this experiment, we assume that we know a lot about crocodiles: they are animals, dangerous, and scaly – and they are crocodiles. Of alligators, we only know that they are alligators. (Note that we do not have to assume that we know all properties of crocodiles. In particular we do not have to assume that we know whether they are alligators.) What we want to know in this experiment is how the distributional similarity of *alligator* and *crocodile*, which in our distributional model is 0.93, will affect the degree of property uncertainty that we have about alligators. In particular, we will measure the probability of sentence φ_1 above, "all alligators are dangerous", according to the prior probability distribution P_0 and the posterior P_1.

To do sampling, we make use of the fact that for a small, finite domain size, it is possible to explicitly generate worlds (Richardson & Domingos 2006, Goodman & Lassiter 2014). We generate a world by probabilistically assigning properties to entities in a fixed-size universe. By repeatedly generating worlds through this generative model, we approximate P_0 through sampling. Generative models are typically described in terms of a *generative story*, a list of instructions on how the relevant data is generated according to the generative model. Often, this generative story is not carried out and is just an abstract characterization of the model. In our case, the generative story is also a high-level description of the program that was used to generate the worlds in practice. The generative story looks like this.

- Fix a domain size n for the universe $\mathcal{U}$. We use $n = 10$.
- Fix a finite collection of predicate symbols. We use *alligator′*, *animal′*, *crocodile′*, *dangerous′*, *scaly′*.
- Fix an extension membership probability p_{ext}. We use $p_{ext} = 0.5$.
- For each predicate symbol $q' \in \{\textit{alligator}', \textit{crocodile}'\}$:

  - For each entity e in the domain: Flip a coin with bias p_{ext} to determine whether e is in the extension of q' or not.

- For each predicate symbol $q' \in \{\textit{animal}', \textit{dangerous}', \textit{scaly}'\}$:
  - For each entity e in the domain: If e is in the extension of *crocodile*′, then e is also in the extension of q'. (This way we make sure that all crocodiles are animals, dangerous, and scaly.)
    If e is not in the extension of *crocodile*′, flip a coin with bias p_{ext} to determine whether e is in the extension of q'.

In all worlds generated in this way, all crocodiles will be animals, dangerous, and scaly, because these subset relations are enforced by the generative process. But the worlds that we generate will differ in whether all alligators are animals, dangerous, or scaly, because the probabilistic process does not enforce any particular relations between these properties. So the agent whose probabilistic information state we model has no uncertainty about the properties of crocodiles (except about whether they are alligators), but is quite uncertain about the properties of alligators. Note that this generative model specifies the degree of uncertainty about properties that the agent has, but does not specify whether the agent has any uncertainty about extensions, and in fact even the extension of *crocodile* is allowed to vary randomly across worlds. This is because our model influences (and is influenced by) uncertainty about properties, but not (at least not directly) uncertainty about extensions.

We can estimate the probability of a sentence φ under the probability distribution P_0 by probabilistically generating worlds with the generative model and counting in how many of the generated worlds φ is true. But because we know how worlds are generated by the generative model, we can analytically determine the probabilities of some sentences of interest under P_0. The probability of "all crocodiles are dangerous" is 1, because the generative process is such that any entity that is a crocodile will also be in the extension of *dangerous*′. Next, we consider the sentence "all alligators are crocodiles." If an entity is an alligator, its probability of being a crocodile as well is $p_{ext} = 0.5$ (as the agent that we are modeling does not know any better), so the probability that any given entity will be both a crocodile and an alligator is $0.5 \cdot 0.5 = 0.25$. The sentence "all alligators are crocodiles" is true in a world if every entity is either not an alligator (which is the case with probability 0.5) or both an alligator and a crocodile (which is the case

with probability 0.25), so because we have 10 entities, the probability of the sentence under P_0 is $(0.75)^{10} = 0.06$. Finally, we look at the sentence we are most interested in, "φ_1: all alligators are dangerous". Its probability under P_0 is actually considerably higher than that of "all alligators are crocodiles", because the process by which we determine whether any given entity is dangerous differs from the process by which we determine whether that entity is a crocodile: We first flip a coin to determine whether it is a crocodile, and if so, it is definitely dangerous; if it is not a crocodile, it will still have a probability of 0.5 of being dangerous. So the probability, for any given entity, of being dangerous is $0.5 \cdot 1 + 0.5 \cdot 0.5 = 0.75$, while the probability for that entity being a crocodile is only 0.5. This is correct, as it reflects the fact that supersets have to be bigger than subsets. So sentence φ_1 is true in a world if every entity is either not an alligator (which is the case with probability 0.5) or is both an alligator and dangerous (which is the case with probability $0.5 \cdot 0.75 = 0.375$), so the probability of the sentence φ_1 under P_0 is $(0.875)^{10} = 0.26$. This is our baseline or chance level probability of sentence φ_1 being true.

We can estimate the probability of a sentence φ under the posterior probability distribution P_1 by sampling worlds according to P_1 and checking, for each of the sampled worlds, whether φ is true in them. To sample worlds according to P_1, we formulate the relevant soft meaning postulates (Equation (19)) as sampling conditions. Say we want to apply a soft meaning postulate involving the extensions of u' and v' in world w. Then we first need to determine o, the observed degree of property overlap of the extensions of u' and v' in w. From o, we can then compute $f(o)$, the fitted value for overlap o, according to the linear regression model from the previous section. This model had an intercept of $\beta_0 = 0.35$ and slope of $\beta_1 = 1.80$. Then we sample a value s' at random from g_o, where g_o is a normal distribution with mean $f(o)$ and a standard deviation which is the residual standard error of the linear regression, in our case 0.15. Finally, we compare the sampled value s' to the observed distributional similarity s of the words u and v. If $s' \geq s$, we add the world w to our sample, otherwise we reject w.

The programming language Church, a language for specifying (small) probabilistic generative models (Goodman, Mansighka, et al. 2008), was used to sample 2,000 worlds according to P_0, and 2,000 worlds according to P_1.[10]

10 The code for this experiment is available at https://utexas.box.com/s/wqt8exy2hlemeh3seggcgr4a1hxnwp5c.

Sentence	words	sim	prior	posterior
φ_1: All alligators are dangerous	alligator, crocodile	0.93	0.26	0.47
φ_2: All alligators are edible	alligator, trout	0.68	0.26	0.38

Table 4 Distributional evidence (similarity *sim* between given *words*) leads to probabilistic inference about properties of alligators: Sampling estimate of the probability of a *sentence*, without distributional evidence (prior) and with distributional evidence (posterior).

The results are shown in the upper row of Table 4. The sentence is φ_1. The vocabulary is the set of properties used in the experiment. The columns *words* and *sim* describe the distributional evidence, in our case, that *alligator* and *crocodile* have a similarity of 0.93. The *prior* column shows the estimated (baseline, or chance-level) probability of φ_1 under P_0, where "estimated" means that we counted the number of sampled worlds in which φ_1 is true. This probability is 0.26, the same as the probability we determined analytically earlier. The posterior column shows the estimated probability of φ_1 under P_1, which is 0.47. The absolute values of the prior and posterior probabilities are not relevant; they depend on the domain size and on the extension membership probability. What is relevant is that the posterior is significantly higher than the prior according to a chi-squared test.[11,12]

The left panel of Figure 5 looks in more detail at how the property overlap between alligators and crocodiles changes when the distributional evidence is introduced. The x-axis lists possible values of property overlap for the extensions of *alligator*′ and *crocodile*′,[13] and the y-axis shows, for each property overlap value, the number of worlds in our sample of 2,000

11 $X^2 = 198.17$, $df = 1$, $p \ll 0.001$, where "$\ll$" means "much smaller than."

12 To test whether there are significantly more worlds in the P_1 sample than in the P_0 sample that make φ_1 true, we consider the null hypothesis that the numbers are the same except for random fluctuations. We construct a probability distribution of values we would expect to see under the null hypothesis. If the actually observed difference in numbers is very unlikely under this distribution, in our case $p \ll 0.001$, we reject the null hypothesis and state that the difference is significant. This is the same hypothesis testing idea that we used in Section 6.6 to determine the probability of a piece of distributional evidence in a given world w.

13 For this data, the denominator of equation (11) is always 5, the number of predicate symbols, because crocodiles are definitely crocodiles, animals, dangerous, and scaly, and alligators are definitely alligators.

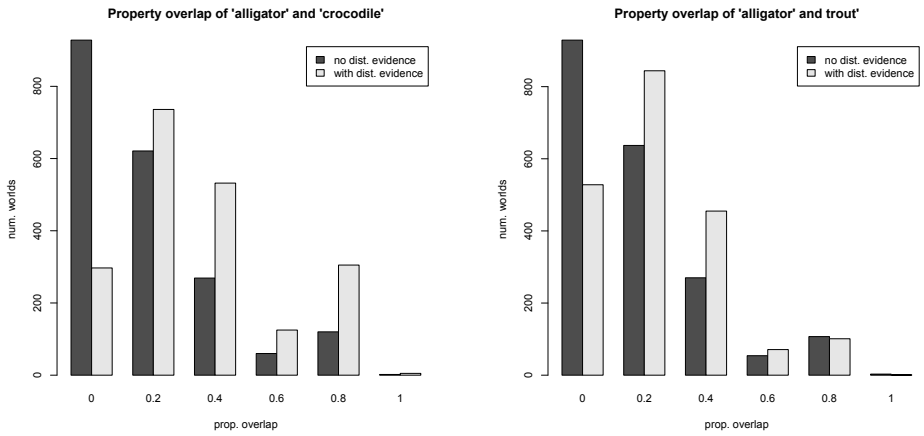


Figure 5 Sampling estimate of the degree of extensional property overlap with *alligator′*, for *crocodile′* (left) and *trout′* (right), without distributional evidence (prior, black) and with distributional evidence (posterior, gray)

that had that value. The bars in black indicate the numbers for the prior P_0, and the gray bars are the numbers for the posterior P_1. As can be seen, the overlap values for P_1 tend to be higher than for P_0. This means that property overlap for *crocodile′* and *alligator′* tends to be higher when the distributional evidence is taken into account. Higher property overlap means that alligators and crocodiles tend to have more of the same properties. So the distributional evidence leads to an inference that probabilistically ascribes to alligators the properties that crocodiles are known to have, including the property of being dangerous. Both Table 4 and Figure 5 confirm that the model supports property inference from distributional data.

Experiment 2: Are alligators edible? The second experiment uses sentence φ_2, "all alligators are edible." We test the influence of the distributional similarity of *alligator* and *trout* on the probability of φ_2. Of trouts, we know that they are animals, aquatic, and edible, as well as trouts. Of alligators, we again only know that they are alligators. We conduct this second experiment to test the influence of different levels of distributional similarity: While the similarity of *alligator* to *crocodile* is 0.93, the similarity of *alligator* and *trout* is only 0.68 in our distributional model.

The generative model is like in the previous experiment, except that the relevant set of properties is different. In analogy to experiment 1, we enforce that all entities that are trouts are also animals, aquatic, and edible. As the number of properties is the same as in experiment 1, all numbers are comparable, except for the difference in distributional similarity.

The second row of Table 4 shows the results of this experiment. The sentence is φ_2, and the similarity of *alligator* and *trout* is 0.68. As expected, the estimated prior probability of φ_2 is 0.26, the same as for φ_1, as all parameters determining the prior probability are the same for experiments 1 and 2. The posterior probability for φ_2 is 0.38, and while this is again significantly higher than the prior,[14] the probability is considerably lower than for φ_1. This is as it should be: We use the strength of the distributional rating as a confidence rating that alligators and trouts share many properties. As this rating is lower than in the case of alligators and crocodiles, the probability that we ascribe to alligators sharing any property of trouts (for instance being edible) should be lower than the probability that we ascribe to alligators sharing any property of crocodiles.

The right panel of Figure 5 shows how property overlap of alligators and trouts changes when the distributional evidence about *alligator/trout* is introduced. The x-axis again lists possible property overlap values for the extensions of *alligator′* and *trout′*, and the y-axis shows, for each property overlap value, the number of sampled worlds that were found to have that value. As in experiment 1, the posterior overlap values (gray) are generally higher than the prior counts (black), though as expected the shift is not as strong as for experiment 1.

Again, the results confirm that the model supports distributional property inference. In addition, they confirm that higher distributional similarity leads to stronger probabilistic inference.

Experiment 3: Are alligators animals? The previous two experiments looked at the effect of adding a single piece of distributional evidence. The third experiment tests what happens when two pieces of distributional evidence are present. Suppose an agent does not know what alligators are, but knows that they are distributionally similar to both crocodiles and trouts, which are both known to be animals. Then the agent should gain even more certainty

14 chi-squared, $X^2 = 60.85$, $df = 1$, $p \ll 0.001$

Sentence	sim of *alligator* to…	prior	posterior
φ_3: All alligators are animals	crocodile: 0.93	0.53	0.68
	trout: 0.68	0.53	0.63
	crocodile: 0.93, trout: 0.68	0.53	0.80

Table 5 Distributional evidence accumulates: Sampling estimate of the probability of a sentence, without distributional evidence (prior) and with distributional evidence (posterior). First two lines: sampling constraint given only for *crocodile*, or only for *trout*. Last line: sampling constraint given for both.

that alligators are probably animals. To test this experimentally, we consider the probability of φ_3, "all alligators are animals".

For this experiment we use a generative model similar to the ones above, but with a larger vocabulary containing *alligator*, *crocodile*, *trout*, *animal*, *aquatic*, *dangerous*, *edible*, and *scaly*. The generative model enforces that all crocodiles are animals, dangerous and scaly, and that all trouts are animals, aquatic, and edible.[15]

As we saw above, enforcing that any crocodile is dangerous increases the probability that an arbitrary entity will be dangerous. As we now enforce that any crocodile or trout must be an animal, the probability under P_0 of an arbitrary entity being an animal is even higher. An entity will be an animal if it is both a crocodile and a trout, a crocodile but not a trout, or a trout but not a crocodile (probability $0.25 + 0.25 + 0.25 = 0.75$), or if it is neither a crocodile nor a trout but still an animal (probability $0.25 \cdot 0.5 = 0.125$), so the probability of an arbitrary entity being an animal is 0.875. Then sentence φ_3 is true in a world if every entity is either not an alligator (which is the case with probability 0.5), or it is both an alligator and an animal (which is the case with probability $0.5 \cdot 0.875 = 0.4375$), so with a domain of size 10, the probability of φ_3 under P_0 is $(0.9375)^{10} = 0.524$.

We sample 2,000 worlds from the prior distribution, and we sample 2,000 worlds each from three different posterior distributions: Distributional

15 The Church implementation for this experiment is available at https://utexas.box.com/s/wqt8exy2hlemeh3seggcgr4a1hxnwp5c.

evidence stating the similarity of *alligator* and *crocodile* (but no information about similarity to *trout*), distributional evidence only for *alligator/trout*, and distributional evidence for both *alligator/crocodile* and *alligator/trout*. The results are shown in Table 5. The estimated probability of φ_3 under the prior probability distribution is shown in the "prior" column. At 0.53, it is reasonably close to the analytically determined value of 0.524. The estimated probabilities of φ_3 under the three different posterior distributions are shown in the "posterior" column: 0.68 when we have distributional evidence about *alligator/crocodile* only, 0.63 with *alligator/trout* only, and 0.8 when we know the similarity of *alligator* to both *crocodile* and *trout*. The posterior probability is significantly higher than the prior in all three conditions.[16] Comparing the first two conditions, we see that as before, the higher distributional similarity of *alligator* to *crocodile* than to *trout* leads to higher certainty of the inference. Crucially, when we compare the third condition to the first two, we see that distributional evidence does indeed accumulate, as the posterior probability, at 0.8, is significantly higher than in the other conditions.[17] So the more types of animals we know that are distributionally similar to *alligator*, the more certain we will be that alligators are animals as well.[18]

These results confirm that the model supports accumulating evidence, where the probabilistic inference about a property q becomes stronger the more distributional evidence the agent has about other terms that have property q.

Overall, the three experiments in this section have shown that the model proposed in this paper lets an agent probabilistically infer properties of an unknown word from distributional evidence. In the prior setting, without distributional evidence, the simulated agent had no knowledge of what prop-

16 Chi-squared tests were again used. For evidence only about *crocodile*, we have: $X^2 = 87.92$, $df = 1$, $p \ll 0.0001$. For evidence only about *trout*: $X^2 = 38.23$, $df = 1$, $p \ll 0.0001$. For evidence about both: $X^2 = 312.28$, $df = 1$, $p \ll 0.0001$. So the differences are significant even with Bonferroni correction for multiple testing.

17 A chi-squared test was performed to compare the condition with distributional information about *crocodile* only to the condition with distributional information about *crocodile* and *trout*. The result was: $X^2 = 72.36$, $df = 1$, $p \ll 0.0001$.

18 Incidentally, the agent will also infer, based on the distributional similarity of *alligator* to *trout*, that alligators may be fish. This is interesting because it shows the interaction between distributional evidence and other knowledge: If the agent knows, say from school, that crocodiles are definitely reptiles and that reptiles and fish are disjoint, then the agent will give nonzero probability to some worlds where all alligators are reptiles but not fish, and to some worlds where all alligators are fish but not reptiles – but zero probability to all worlds where alligators are both reptiles and fish.

erties alligators might have, besides being alligators. When distributional evidence was taken into account, the probabilistic information state showed a considerably higher likelihood that alligators are dangerous, edible, and animals (respectively, in the three experiments). The simulated agent was able to make these probabilistic inferences without explicitly learning about either the properties or the extension of *alligator*. In more detail, experiments 1 and 2 have shown that the stronger the distributional evidence, that is, the higher the distributional similarity between *alligator* and a word u, the stronger the probabilistic inference that alligators have the properties that u is known to have. Experiment 3 has shown that the model accumulates evidence from distributional similarity to multiple words.

8 Related work: Distributional information and formal semantics

In this section, I discuss existing papers that link distributional models to either logical form as it is used in computational linguistics, or to semantic theory. There are not many papers that propose such a link, and the papers that exist differ widely in their backgrounds, their goals, and in the proposed models. So my aim in this section is not to unify all those papers into a single story, but to showcase the current variety in people's thinking about the topic.

In computational linguistics, Beltagy et al. 2013 and Lewis & Steedman 2013 extend logical form representations of sentence meaning with distributional information. Their aim is to address polysemy and to obtain additional inference rules beyond those that can be extracted from manually created lexical resources. In particular the approach that we proposed in Beltagy et al. is related to this paper. In Beltagy et al., we computed the probabilities of truth assignments based on the weights of distributional evidence. But we used a heuristic to get from distributional weights to probabilities of truth assignments, while the current paper proposes a principled mechanism for the influence of distributional evidence on the probability of a world. Lewis & Steedman 2013 uses distributional information to induce word senses and to decide on the preferred sense in a given context, but at the sentence level they use standard first-order logic.

The underlying question in Erk 2013 is how inferential information encoded distributionally could be linked to a model-theoretic semantics. I suggested that distributional representations are modeling mental concepts, as they have successfully been used to model different phenomena connected

to conceptual representations. I further suggested that intensions are linked to mental concepts, providing a connection between distributional representations and model theory. But to make this possible, the paper had to disregard gradedness, one of the core properties of distributional models. Also, the nature of the connection between distributional representations, mental concepts, and intensions was not worked out.

The main aim of Copestake & Herbelot 2013 is to point out a formal connection between model theory and distributional semantics, namely through the sentences that are true of an entity in the world. They postulate an "ideal distribution" as a distributional model computed from a corpus that consists of logical forms of all true statements about all entities in the world. The actual distribution that a speaker perceives is then an approximation of the ideal distribution. The approach is elegant in that it reduces distributional information to information on what is true about entities in the world. It is also interesting in its change of focus from the words that make up the corpus to the entities that the corpus is about; Herbelot continues the work on individuals and their distributional trace in Herbelot 2015. But the reliance on ideal distributions is a drawback of the approach. Actual distributions do not approximate ideal distributions too well. They are noisy and oddly eclectic. So I feel that it is important to discuss, as I did in this paper, how a speaker can learn from distributional data that is imperfect in all these ways, and also how noisy distributional information can be integrated with other knowledge.

Larsson (2015) is interested in learning from perceptual information. He uses vector space representations, but he uses them not for distributional but low-level perceptual information. He defines perceptual meanings as perceptron classifiers operating on the perceptual information, and views these classifiers as intensions. Larsson discusses the connection of his work to distributional approaches, in particular whether his perceptual vectors could be viewed as distributional. They could: There are distributional models that use perceptual or mixed perceptual and textual dimensions (Andrews, Vigliocco & Vinson 2009, Feng & Lapata 2010, Bruni et al. 2012). Cooper et al. (2014) take the idea further by applying perceptual classifiers to situations, which are abstract objects with associated features that are propositions. As discussed in Section 5, they frame this idea in a probabilistic system of types, where situations are assigned types that are propositions, and this assignment is associated with probabilities.

It remains to be seen if in the future there will be convergence on the question of how distributional and formal semantics should be linked, and to what end.

9 Conclusion

In this paper I have suggested that (for suitable distributional models) distributional similarity signals property overlap. This is possible because distributional evidence keeps track of the predicates that apply to a noun target, and the selectional constraints of a predicate indicate properties of its argument. Because of this, distributional evidence can be used for property inference: "I don't know what an alligator is, but it must be something like a crocodile, so it is probably an animal."

I have also proposed a mechanism for distributional property inference. It assumes that an agent has a probabilistic information state that can make use of noisy, incomplete data like distributional data. Based on the observation that *alligator* and *crocodile* have a high distributional similarity, the agent infers that the two words must have a high property overlap, and adopts a probabilistic information state that ascribes higher probability to worlds in which this is the case. The mechanism focuses on an agent's uncertainty about properties, not their uncertainty about extensions. I have pointed out that an agent can be highly certain that alligators are animals without being certain about the extension of *alligator′*, which explains how an agent can use a word felicitously without being aware of its extension.

This paper takes a first step in the direction of integrating formal semantics with distributional information, but at this point there are of course many more open questions than solved ones. On the topic of learning word meaning from distributional evidence, one obvious question is what happens when properties only apply to some but not all entities in an extension. If alligators are distributionally similar to crocodiles, then can we draw any conclusions from the fact that some but not all crocodiles live in zoos? Also, not all of word meaning can be easily described through verbalizable properties. For example, what distinguishes the word *skittish* from *nervous*? Andrews, Vigliocco & Vinson (2009) suggest that it is words like *crime*, *justice*, and *finance* where distributional evidence is most important to word learning. But if the meaning of such words cannot easily be defined through property lists, then what is an agent learning about them from distributional data? Another question concerns learning from data that is not distributional. For

example, what if an agent has once seen an alligator in the zoo, and has also acquired distributional information about these creatures: How do the two knowledge sources combine?

Polysemy is another big issue to address, learning the meanings of polysemous words, but also determining the meaning of a polysemous word in a particular sentence context. It has been suggested that distributional information can help on this task (Erk 2010, McNally & Boleda 2014), and maybe the mechanism from this paper can prove useful for addressing polysemy as well. A probabilistic framework, like the one introduced in this paper, should be useful in characterizing word meaning in context, as differences in meaning seem to come in degrees (Brown 2008, Erk, McCarthy & Gaylord 2013). And maybe the task of describing what a word means in a particular sentence can be framed as property inference, namely inferring the most relevant properties of a word in a given context.

Another issue that needs further work is the probabilistic inference framework. As discussed in Section 5, the probabilistic logic that we use is problematic in that it is currently not clear how to extend it to infinitely many worlds. One way to address this is to find a way to extend it, a direction that Clarke & Keller (2015) take; another is to pursue a formulation in a situation-based framework like that of Cooper et al. (2014) or Goodman & Lassiter (2014).

Finally, there is a need for more research on distributional models for property inference, to develop efficient models beyond the initial approaches proposed by Johns & Jones (2012), Făgărăşan, Vecchi & Clark (2015), Herbelot & Vecchi (2015) and Gupta et al. (2015) and to see what kinds of properties can be reliably learned and whether verb properties can be learned as well as noun properties.

The aim of this paper was to argue for distributional similarity as property overlap, and to provide a first simple model for how agents may use distributional evidence in a probabilistic framework. I hope that this model will prove to be extensible to addressing some of these open questions in the future.

References

Agirre, Eneko, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca & Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *North American Chapter of the Association for Computational Linguistics: Human language technologies*

(NAACL HLT). Boulder, CO. http://aclweb.org/anthology/N/N09/N09-1003.pdf.

Almuhareb, Abdulrahman & Massimo Poesio. 2004. Attribute-based and value-based clustering: An evaluation. In *Conference on empirical methods in natural language processing (EMNLP)*, 158–165. http://aclweb.org/anthology/W/W04/W04-3221.pdf.

Almuhareb, Abdulrahman & Massimo Poesio. 2005. Finding concept attributes in the web using a parser. In *Proceedings of the corpus linguistics conference*. Birmingham. http://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2005-journal/Thewebasacorpus/cl133papalmuhareb.pdf.

Andrews, Mark, Gabriella Vigliocco & David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review* 116(3). 463–498. http://dx.doi.org/10.1037/a0016261. http://www.mjandrews.net/papers/andrews.psychrev.2009.pdf.

Baroni, Marco, Raffaella Bernardi & Roberto Zamparelli. 2014. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology* 9(6). 5–110. http://elanguage.net/journals/lilt/article/view/3746.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3). 209–226. http://dx.doi.org/10.1007/s10579-009-9081-4. http://clic.cimec.unitn.it/marco/publications/wacky-lrej.pdf.

Baroni, Marco, Georgiana Dinu & Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *52nd annual meeting of the Association for Computational Linguistics (ACL)*. http://aclweb.org/anthology/P/P14/P14-1023.pdf.

Baroni, Marco & Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics* 36(4). 673–721. http://aclweb.org/anthology/J/J10/J10-4006.pdf.

Baroni, Marco & Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Workshop on geometrical models of natural language semantics (GEMS)*. Edinburgh, Great Britain. http://aclweb.org/anthology/W/W11/W11-2501.pdf.

Baroni, Marco, Brian Murphy, Eduard Barbu & Massimo Poesio. 2010. Strudel: A corpus-based semantic model based on properties and types. *Cognitive*

Science 34(2). 222–254. http://clic.cimec.unitn.it/strudel/materials/strudel.pdf.

Baroni, Marco & Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Conference on empirical methods in natural language processing (EMNLP)*. Cambridge, MA. http://aclweb.org/anthology/D/D10/D10-1115.pdf.

Beltagy, Islam, Cuong Chau, Gemma Boleda, Dan Garrette, Katrin Erk & Raymond Mooney. 2013. Montague meets Markov: Deep semantics with probabilistic logical form. In *Second joint conference on lexical and computational semantics (*SEM)*. Atlanta, GA. http://aclweb.org/anthology/S/S13/S13-1002.pdf.

van Benthem, Johan, Jelle Gerbrandy & Barteld Kooi. 2009. Dynamic update with probabilities. *Studia Logica* 93(1). 67–96. http://dx.doi.org/10.1007/s11225-009-9209-y. http://www.gerbrandy.com/science/papers/PP-2006-21.text.pdf.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning (information science and statistics)*. Secaucus, NJ: Springer-Verlag New York.

Blei, David, Andrew Ng & Michael Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3(4-5). 993–1022. http://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf.

Brown, Susan W. 2008. Choosing sense distinctions for WSD: Psycholinguistic evidence. In *46th annual meeting of the Association for Computational Linguistics: Human language technologies (ACL HLT)*. Columbus, OH. http://aclweb.org/anthology/P/P08/P08-2063.pdf.

Bruni, Elia, Gemma Boleda, Marco Baroni & Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *50th annual meeting of the Association for Computational Linguistics (ACL)*. Jeju Island, Korea. http://aclweb.org/anthology/P/P12/P12-1015.pdf.

Clarke, Daoud & Bill Keller. 2015. Efficiency in ambiguity: Two models of probabilistic semantics for natural language. In *11th international conference on computational semantics (IWCS)*. London, Great Britain. http://aclweb.org/anthology/W/W15/W15-0118.pdf.

Coecke, Bob, Mehrnoosh Sadrzadeh & Stephen Clark. 2011. Mathematical foundations for a compositional distributed model of meaning. *Linguistic Analysis* 36(1-4). A Festschrift for Joachim Lambek, 345–384. http://www.cs.ox.ac.uk/files/2879/LambekFestPlain.pdf.

Collobert, Ronan & Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *25th*

international conference on machine learning (ICML). Helsinki, Finland. http://ronan.collobert.com/pub/matos/2008_nlp_icml.pdf.

Cooper, Robin, Simon Dobnik, Shalom Lappin & Staffan Larsson. 2014. A probabilistic rich type theory for semantic interpretation. In *EACL 2014 workshop on type theory and natural language semantics (TTNLS)*. Gothenburg, Sweden. http://aclweb.org/anthology/W/W14/W14-1409.pdf.

Copestake, Ann & Aurélie Herbelot. 2013. Lexicalised compositionality. Unpublished draft. http://www.cl.cam.ac.uk/~aac10/papers/lc3-oweb.pdf.

Cruse, Alan. 2000. *Meaning in language: An introduction to semantics and pragmatics.* Oxford University Press.

Curran, James, Stephen Clark & Johan Bos. 2007. Linguistically motivated large-scale NLP with C&C and Boxer. In *45th annual meeting of the Association for Computational Linguistics companion volume: Demo and poster sessions*, 33–36. Prague, Czech Republic. http://aclweb.org/anthology/P/P07/P07-2009.pdf.

van Deemter, Kees. 2013. The two cultures of logic. In *On fuzziness*, 719–724. Berlin & Heidelberg, Germany: Springer. http://homepages.abdn.ac.uk/k.vdeemter/pages/for-zadeh.pdf.

Devereux, Barry, Nicholas Pilkington, Thierry Poibeau & Anna Korhonen. 2009. Towards unrestricted, large-scale acquisition of feature-based conceptual representations from corpus data. *Research on Language and Computation* 7. 137–170.

Edmonds, Philip & Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics* 28(2). 105–144. http://dx.doi.org/10.1162/089120102760173625. http://www.cs.toronto.edu/pub/gh/Edmonds+Hirst-2002.pdf.

van Eijck, Jan & Shalom Lappin. 2012. Probabilistic semantics for natural language. In Zoe Christoff, Paulo Galeazzi, Nina Gierasimczuk, Alexandru Marcoci & Sonja Smets (eds.), *Logic and interactive rationality (LIRA) yearbook*, vol. 2, 17–35. Amsterdam dynamics group. http://homepages.cwi.nl/~jve/papers/13/pdfs/vaneijcklappinLIRA.pdf.

Erk, Katrin. 2010. What is word meaning, really? (And how can distributional models help us describe it?) In *Workshop on geometrical models of natural language semantics (GEMS)*. Uppsala, Sweden. http://aclweb.org/anthology/W/W10/W10-2803.pdf.

Erk, Katrin. 2013. Towards a semantics for distributional representations. In *10th international conference on computational semantics (IWCS)*. Potsdam, Germany. http://aclweb.org/anthology/W/W13/W13-0109.pdf.

Erk, Katrin, Diana McCarthy & Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics* 39(3). 511–554. http://dx.doi.org/10.1162/COLI_a_00142.

Făgărăşan, Luana, Eva Maria Vecchi & Stephen Clark. 2015. From distributional semantics to feature norms: Grounding semantic models in human perceptual data. In *11th international conference on computational semantics (IWCS)*. London, Great Britain. http://aclweb.org/anthology/W/W15/W15-0107.pdf.

Fellbaum, Christiane (ed.). 1998. *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Feng, Yansong & Mirella Lapata. 2010. Visual information in semantic representation. In *Human language technologies: The 11th annual conference of the North American chapter of the Association for Computational Linguistics (HLT NAACL)*. Los Angeles, CA. http://aclweb.org/anthology/N/N10/N10-1011.pdf.

Fine, Kit. 1975. Vagueness, truth and logic. *Synthese* 30(3/4). 265–300. http://www.niu.edu/~gpynn/Fine_Vagueness_Truth&Logic.pdf.

Fu, Ruiji, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang & Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *52nd annual meeting of the Association for Computational Linguistics (ACL)*. Baltimore, MD.

Goodman, Noah D. & Daniel Lassiter. 2014. Probabilistic semantics and pragmatics: Uncertainty in language and thought. In Shalom Lappin & Chris Fox (eds.), *Handbook of contemporary semantics*. Wiley-Blackwell. http://www.stanford.edu/~ngoodman/papers/Goodman-HCS-final.pdf.

Goodman, Noah D., Vihash K. Mansighka, Daniel Roy, Keith Bonawitz & Joshua B. Tenenbaum. 2008. Church: A language for generative models. In *Uncertainty in artificial intelligence*. Helsinki, Finland. http://stanford.edu/~ngoodman/papers/churchUAI08_rev2.pdf.

Goodman, Noah D., Joshua B. Tenenbaum & Tobias Gerstenberg. to appear. Concepts in a probabilistic language of thought. In Eric Margolis & Stephen Laurence (eds.), *The conceptual mind: New directions in the study of concepts*. MIT Press. http://www.stanford.edu/~ngoodman/papers/ConceptsChapter-final.pdf.

Grefenstette, Edward & Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Conference on empirical methods in natural language processing (EMNLP)*.

Edinburgh, Great Britain. http://aclweb.org/anthology/D/D11/D11-1129.pdf.

Gupta, Abhijeet, Gemma Boleda, Marco Baroni & Sebastian Pado. 2015. Distributional vectors encode referential attributes. In *Conference on empirical methods in natural language processing (EMNLP)*. Lisbon, Portugal. http://aclweb.org/anthology/D/D15/D15-1002.pdf.

Herbelot, Aurélie. 2015. Mr Darcy and Mr Toad, gentlemen: Distributional names and their kinds. In *11th international conference on computational semantics (IWCS)*. London, Great Britain. http://aclweb.org/anthology/W/W15/W15-0120.pdf.

Herbelot, Aurélie & Eva Maria Vecchi. 2015. Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Conference on empirical methods in natural language processing (EMNLP)*. Lisbon, Portugal. http://aclweb.org/anthology/D/D15/D15-1003.pdf.

Jaccard, Paul. 1901. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles* 37. 241–272. http://www.researchgate.net/publication/243457811_Distribution_de_la_flore_alpine_dans_le_bassin_des_Dranses_et_dans_quelques_rgions_voisines.

Johns, Brendan T & Michael N Jones. 2012. Perceptual inference through global lexical similarity. *Topics in Cognitive Science* 4(1). 103–120. http://www.indiana.edu/~clcl/Papers/JohnsJones_TopiCS.pdf.

Koehn, Philipp & Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *ACL workshop on unsupervised lexical acquisition*. http://homepages.inf.ed.ac.uk/pkoehn/publications/learnlex2002.pdf.

Kotlerman, Lili, Ido Dagan, Idan Szpektor & Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering* 16(4). 359–389. http://dx.doi.org/10.1017/S1351324910000124. http://eprints.pascal-network.org/archive/00008618/01/directional-distsim.pdf.

Kremer, Gerhard & Marco Baroni. 2010. Predicting cognitively salient modifiers of the constitutive parts of concepts. In *2010 workshop on cognitive modeling and computational linguistics*. Uppsala, Sweden. http://aclweb.org/anthology/W/W10/W10-2007.pdf.

Landauer, Thomas & Susan 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. http:

//dx.doi.org/10.1037/0033-295X.104.2.211. http://www.stat.cmu.edu/~cshalizi/350/2008/readings/Landauer-Dumais.pdf.

Larsson, Staffan. 2015. Formal semantics for perceptual classification. *Journal of logic and computation* 25(2). 335–369. http://dx.doi.org/10.1093/logcom/ext059. http://www.ling.gu.se/~sl/Papers/jlc-preprint.pdf.

Lazaridou, Angeliki, Elia Bruni & Marco Baroni. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *52nd annual meeting of the Association for Computational Linguistics (ACL)*. Baltimore, MD. http://aclweb.org/anthology/P/P14/P14-1132.pdf.

Lenci, Alessandro. 2008. Distributional approaches in linguistic and cognitive research. *Italian Journal of Linguistics* 20(1). 1–31. http://linguistica.sns.it/RdL/20.1/ALenci.pdf.

Lenci, Alessandro & Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *First joint conference on lexical and computational semantics (*SEM)*. Montréal, Canada. http://aclweb.org/anthology/S/S12/S12-1012.pdf.

Levy, Elena & Katherine Nelson. 1994. Words in discourse: A dialectical approach to the acquisition of meaning and use. *Journal of Child Language* 21(2). 367–389. http://dx.doi.org/10.1017/S0305000900009314.

Levy, Omer, Steffen Remus, Chris Biemann & Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Conference of the North American chapter of the Association for Computational Linguistics: Human language technologies (HLT NAACL)*. Denver, CO. http://aclweb.org/anthology/N/N15/N15-1098.pdf.

Lewis, Mike & Mark Steedman. 2013. Combined distributional and logical semantics. *Transactions of the Association for Computational Linguistics (TACL)* 1. 179–192. http://aclweb.org/anthology/Q/Q13/Q13-1015.pdf.

Lin, Dekang. 1998a. An information-theoretic definition of similarity. In *15th international conference on machine learning (ICML)*, 296–304. http://dl.acm.org/citation.cfm?id=657297.

Lin, Dekang. 1998b. Automatic retrieval and clustering of similar words. In *36th annual meeting of the Association for Computational Linguistics and the 17th international conference on computational linguistics (COLING ACL)*. Montreal, Canada. http://aclweb.org/anthology/P/P98/P98-2127.pdf.

Lin, Dekang & Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering* 7(4). 343–360. http://dx.doi.

org/10.1017/S1351324901002765. http://cluster.cis.drexel.edu:8080/sofia/resources/QA.Data/PDF/2001_NLEng_Lin_and_Pantel_Discovery_of_Inference_Rules_for_Question_Answering-2826639643/2001_NLEng_Lin_and_Pantel_Discovery_of_Inference_Rules_for_Question_Answering.pdf.

McDonald, Scott & Michael Ramscar. 2001. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Cognitive Science Society (CogSci)*, 611–616. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.7535&type=pdf.

McNally, Louise & Gemma Boleda. 2014. Conceptual vs. referential affordance in concept composition. unpublished draft. http://www.upf.edu/pdi/louise-mcnally/_pdf/publications/McNally-Boleda-14.pdf.

McRae, Ken, George S. Cree, Mark S. Seidenberg & Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods* 37(4). 547–559. http://dx.doi.org/10.3758/BF03192726. https://sites.google.com/site/kenmcraelab/publications/McRae_etal_norms_BRM_05.pdf?attredirects=0.

Mikolov, Tomas, Martin Karafiát, Lukáš Burget, Jan Černocký & Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, 2877–2880. http://www.fit.vutbr.cz/research/groups/speech/publi/2010/mikolov_interspeech2010_IS100722.pdf.

Mitchell, Jeff & Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science* 34(8). 1388–1429. http://dx.doi.org/10.1111/j.1551-6709.2010.01106.x.

Murphy, Greg L. 2002. *The big book of concepts*. MIT Press.

Nilsson, Nils J. 1986. Probabilistic logic. *Artificial intelligence* 28(1). 71–87. http://dx.doi.org/10.1016/0004-3702(86)90031-7. http://ai.stanford.edu/~nilsson/OnlinePubs-Nils/PublishedPapers/problogic.pdf.

Padó, Sebastian, Ulrike Padó & Katrin Erk. 2007. Flexible, corpus-based modelling of human plausibility judgements. In *Conference on empirical methods in natural language processing and the conference on computational natural language learning (EMNLP CoNLL)*. Prague, Czech Republic. http://aclweb.org/anthology/D/D07/D07-1042.pdf.

Peirsman, Yves. 2008. Word space models of semantic similarity and relatedness. In *European summer school in logic, language and information (ESSLLI) student session*. Hamburg, Germany. http://wwwling.arts.kuleuven.be/qlvl/prints/peirsman_2008draft_Word_Space_Models.pdf.

Putnam, Hilary. 1973. Meaning and reference. *The Journal of Philosophy* 70(19). 699–711. http://dx.doi.org/10.2307/2025079. http://home.sandiego.edu/~baber/analytic/Putnam1973.pdf.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/.

Richardson, Matthew & Pedro Domingos. 2006. Markov logic networks. *Machine Learning* 62(1-2). 107–136. http://dx.doi.org/10.1007/s10994-006-5833-1. http://homes.cs.washington.edu/~pedrod/kbmn.pdf.

Roller, Stephen, Katrin Erk & Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *25th international conference on computational linguistics (COLING)*. Dublin, Ireland. http://aclweb.org/anthology/C/C14/C14-1097.pdf.

Sahlgren, Magnus. 2006. *The word-space model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Stockholm University dissertation. https://www.sics.se/~mange/TheWordSpaceModel.pdf.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1). 97–123. http://aclweb.org/anthology/J/J98/J98-1004.pdf.

Snow, Rion, Daniel Jurafsky & Andrew Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *21st international conference on computational linguistics and 44th annual meeting of the Association for Computational Linguistics (COLING ACL)*. Sydney, Australia. http://aclweb.org/anthology/P/P06/P06-1101.pdf.

Socher, Richard, Brody Huval, Christopher D Manning & Andrew Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Joint meeting of the conference on empirical methods in natural language processing and the conference on computational natural language learning (EMNLP CoNLL)*. Jeju Island, Korea. http://aclweb.org/anthology/D/D12/D12-1110.pdf.

Thill, Serge, Sebastian Pado & Tom Ziemke. 2014. On the importance of a rich embodiment in the grounding of concepts: Perspectives from embodied cognitive science and computational linguistics. *Topics in Cognitive Science* 6(3). 545–558. http://dx.doi.org/10.1111/tops.12093.

Turney, Peter & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37(1). 141–188. http://www.jair.org/media/2934/live-2934-4846-jair.pdf.

Veltman, Frank. 1996. Defaults in update semantics. *Journal of Philosophical Logic* 25(3). 221–261. http://dx.doi.org/10.1007/BF00248150. http://staff.science.uva.nl/~veltman/papers/FVeltman-dius.pdf.

Vigliocco, Gabriella, David Vinson, William Lewis & Merrill Garrett. 2004. Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology* 48(4). 422–488.

Wu, Stephen & William Schuler. 2011. Structured composition of semantic vectors. In *9th international conference on computational semantics (IWCS)*. Oxford, Great Britain. http://aclweb.org/anthology/W/W11/W11-0131.pdf.

Zadeh, Lotfi A. 1965. Fuzzy sets. *Information and Control* 8(3). 338–353.

Zeevat, Henk. 2013. Implicit probabilities in update semantics. In Maria Aloni, Michael Franke & Floris Roelofsen (eds.), *Festschrift for Jeroen Groenendijk, Martin Stokhof, and Frank Veltman*. Amsterdam, The Netherlands: ILLC. http://www.illc.uva.nl/Festschrift-JMF/papers/39_Zeevat.pdf.

Katrin Erk
University of Texas at Austin
Linguistics Department
4.734 Liberal Arts Building
305 E 23rd ST B5100
Austin, TX, USA 78712
katrin.erk@mail.utexas.edu