# Towards a semantics for distributional representations

Katrin Erk
University of Texas at Austin
katrin.erk@mail.utexas.edu

**Abstract**

Distributional representations have recently been proposed as a general-purpose representation of natural language meaning, to replace logical form. There is, however, one important difference between logical and distributional representations: Logical languages have a clear semantics, while distributional representations do not. In this paper, we propose a semantics for distributional representations that links points in vector space to mental concepts. We extend this framework to a joint semantics of logic and distributions by linking intensions of logical expressions to mental concepts.

## 1 Introduction

Distributional similarity can model a surprising range of phenomena (e.g., Lund et al. (1995); Landauer and Dumais (1997)) and is useful in many NLP tasks (Turney and Pantel, 2010). Recently, it has been suggested that a general-purpose framework for representing natural language semantics should be distributional, such that it could represent word similarity and phrase similarity (Coecke et al., 2010; Baroni and Zamparelli, 2010; Grefenstette and Sadrzadeh, 2011; Clarke, 2012). Another suggestion has been to combine distributional representations and logical form, with the argument that the strengths of the two frameworks are in complementary areas (Garrette et al., 2011).

One important difference between logic and distributional representations is that logics have a semantics. For example, a model[1] in model-theoretic semantics provides a truth assignment to each sentence of a logical language. More generally, it associates expressions of a logic with set-theoretic structures, for example the constant $cat'$ could be interpreted as the set of all cats in a given world. But what is the interpretation of a distributional representation? What does a point in vector space, where the dimensions are typically uninterpretable symbols, stand for? [2] In this paper, we propose a semantics in which distributional representations stand for mental concepts, and are linked to intensions of logical expressions. This gives us a joint semantics for distributional and logical representations.

*Distributional representations stand for mental concepts.* One central function of models is that they evaluate sentences of a logic as being either true or false. Distributional representations have been evaluated on a variety of phenomena connected to human concept representation (e.g., Lund et al. (1995); Landauer and Dumais (1997); Burgess and Lund (1997)). Here, evaluation means that predictions based on distributional similarity are compared to experimental results from human subjects. So we will interpret distributional representations over a conceptual structure.

*Distributional representations stand for intensions.* Gärdenfors (2004) suggests that the intensions of logical expressions should be mental concepts. By adopting this view, we can link distributional representations and logic through a common semantics: Both the intensions of logical expressions and the interpretations of distributional representations are mental concepts. However, there is a technical

[1] In the context of logical languages, "models" are structures that provide interpretations. In the context of distributional approaches, "distributional models" are particular choices of parameters. To avoid confusion, this paper will reserve the term "model" for the model-of-a-logic sense.

[2] Clark et al. (2008) encode a model in a vector space in which natural language sentences are mapped to a single-dimensional space that encodes truth and falsehood. This is a vector space representation, but it is not distributional as it is not derived from observed contexts. In particular, it does not constitute a semantics for a distributional representation.

$$\frac{\exists x\big(woodchuck(x) \wedge see(John, x)\big) \quad sim(woodchuck, groundhog) > \theta}{\exists x\big(groundhog(x) \wedge see(John, x)\big)}$$

Figure 1: Sketch of an example interaction of distributional and logical representations

problem: If intensions are mental concepts, they cannot be mappings from possible worlds to extensions, which is the prevalent way of defining intensions. We address this problem through *hyper-intensional semantics.* Hyper-intensional approaches in formal semantics (Fox and Lappin, 2001, 2005; Muskens, 2007) were originally introduced to address problems in the granularity of intensions. Crucially, some hyper-intensional approaches have intensions that are abstract objects, with minimal requirements on the nature of these objects. So we can build on them to link some intensions to conceptual structure.

Why design a semantics for distributional representations? Our aim is not to explicitly construct conceptual models; that would be at least as hard as constructing an ontology. Rather, our aim is to support inferences. Distributional representations induce synonyms and paraphrases automatically based on distributional similarity (Lin, 1998; Lin and Pantel, 2001). As Garrette et al. (2011) point out, and as illustrated in Figure 1, these can be used as inference rules within logical form. But when is such inference projection valid? Our main aim for constructing a joint semantics is to provide a principled basis for answering this question.

In the current paper, we construct a first semantics along the lines sketched above. In order to be able to take this first step, we simplify distributional predictions greatly by discretizing them. We want to stress, however, that this is a temporary restriction; our eventual aim is to make use of the ability of distributional models to handle graded and uncertain information as well as ambiguity.

## 2 Related work

**Predicting sentence similarity with distributional representations.** The distributional representation for a word is typically based on the textual contexts in which it has been observed (Turney and Pantel, 2010). The distributional representation of a document is typically based on the words that it contains, or on latent classes derived from co-occurrences of those words (Landauer and Dumais, 1997; Blei et al., 2003). Phrases and sentences occupy an unhappy middle ground between words and documents. They re-appear too rarely for a representation in terms of the textual contexts in which they have been observed, and they are too short for a document-like representation. There are multiple approaches to predicting similarity between sentences based on distributional information. The first computes a single vector space representation for a phrase or sentence in a compositional manner from the representations of the individual words (Coecke et al., 2010; Baroni and Zamparelli, 2010; Grefenstette and Sadrzadeh, 2011). This approach currently still faces big hurdles, including the problem of encoding the meaning of function words and the problem of predicting similarity for sentences of different structure. The second approach compares two phrases or sentences by computing multiple pairwise similarity values between components (words or smaller phrases) of the two sentences and then combining those similarity values (Socher et al., 2011; Turney, 2012). The third approach seeks to transform the representation of one sentence into another through term rewriting, where the rewriting rules are based on distributional similarity between words and smaller phrases (Bar-Haim et al., 2007). The approach of Garrette et al. (2011) can be viewed as falling into the third group. It represents sentences not as syntactic graphs as Bar-Haim et al. (2007) but through logic, and injects weighted inference rules derived from distributional similarity. Our approach belongs into this third group. The aim of the semantics that we present in Section 3 is to show that the use of distributional rewriting rules does not change the semantics of a logical expression. A fourth approach is the taken by Clarke (2007, 2012), who formalizes the idea of "meaning as context" in an algebraic framework that replaces concrete corpora with a generative corpus model that can assign probabilities to arbitrary word sequences. This eliminates the sparseness problem of finite corpora, such that both words and arbitrary phrases can be given distributional representations. Clarke also combines vector spaces and logic-based semantics by proposing a space in which the dimensions

| | |
|---|---|
| (IHTT1) $p \vdash \top$<br>(IHTT2) $\bot \vdash p$<br>(IHTT3) $\vdash \neg p \leftrightarrow p \rightarrow \bot$<br>(IHTT4) $r \vdash p \wedge q$ iff $r \vdash p$ and $r \vdash q$<br>(IHTT5) $p \vee q \vdash r$ iff $p \vdash r$ or $q \vdash r$<br>(IHTT6) $p \vdash q \rightarrow r$ iff $p \wedge q \vdash r$<br>(IHTT7) $p \vdash \forall x_B \phi_{\langle B,\Pi\rangle}$ iff $p \vdash \phi$<br>(IHTT8) $\phi(a) \vdash \exists x_B \phi(x)$ (where $\phi \in \langle B, \Pi\rangle$, and $a$ is a constant in $B$) | (IHTT9) $\vdash \lambda u \phi(v) \cong \phi[u/v]$ (where $u$ is a variable in $A$, $v \in A$, $\phi \in \langle A, B\rangle$, and $v$ is not bound when substituted for $u$ in $\phi$)<br>(IHTT10) $\vdash \forall s, t_\Pi \big(s \cong t \leftrightarrow (s \leftrightarrow t)\big)$<br>(IHTT11) $\vdash \forall \phi, \psi_{\langle A,B\rangle} \big(\forall u_A (\phi(u) \cong \psi(u)) \rightarrow \phi \cong \psi\big)$<br>(IHTT12) $\vdash \forall u, v_A \forall \phi_{\langle A,B\rangle} \big(u = v \rightarrow \phi(u) \cong \phi(v)\big)$<br>(IHTT13) $\vdash \forall t_\Pi \big(t \vee \neg t\big)$ |

Table 1: Axioms of the intensional higher-order type theory IHTT of Fox and Lappin (2001)

| | |
|---|---|
| • If $\alpha_A$ is a non-logical constant, then $\|\|\alpha\|\|^{M,g} = F(I(\alpha))$<br>• If $\alpha_A$ is a variable, then $\|\|\alpha\|\|^{M,g} = g(\alpha)$<br>• $\|\|\alpha_{\langle A,B\rangle}(\beta_A)\|\|^{M,g} = \|\|\alpha\|\|^{M,g}\big(\|\|\beta\|\|^{M,g}\big)$<br>• If $\alpha$ is in $A$ and $u$ is a variable in $B$, then $\|\|\lambda u \alpha\|\|^{M,g}$ is a function $h : D_A \rightarrow D_B$ such that for any $a \in D_A$, $h(a) = \|\|\alpha\|\|^{M,g[u/a]}$<br>• $\|\|\neg\phi_\Pi\|\|^{M,g} = t$ iff $\|\|\phi\|\|^{M,g} = f$<br>• $\|\|\phi_\Pi \wedge \psi_\Pi\|\|^{M,g} = t$ iff $\|\|\phi\|\|^{M,g} = \|\|\psi\|\|^{M,g} = t$ | • $\|\|\phi_\Pi \vee \psi_\Pi\|\|^{M,g} = t$ iff $\|\|\phi\|\|^{M,g} = t$ or $\|\|\psi\|\|^{M,g} = t$<br>• $\|\|\phi_\Pi \rightarrow \psi_\Pi\|\|^{M,g} = t$ iff $\|\|\phi\|\|^{M,g} = f$ or $\|\|\psi\|\|^{M,g} = t$<br>• $\|\|\phi_\Pi \leftrightarrow \psi_\Pi\|\|^{M,g} = t$ iff $\|\|\phi\|\|^{M,g} = \|\|\psi\|\|^{M,g}$<br>• $\|\|\alpha_A \cong \beta_A\|\|^{M,g} = t$ iff $\|\|\alpha\|\|^{M,g} = \|\|\beta\|\|^{M,g}$<br>• $\|\|\alpha_A = \beta_A\|\|^{M,g} = t$ iff $I(\alpha) = I(\beta)$<br>• $\|\|\forall u_A \phi_\Pi\|\|^{M,g} = t$ iff for all $a \in D_A\big(\|\|\phi\|\|^{M,g[u/a]} = t\big)$<br>• $\|\|\exists u_A \phi_\Pi\|\|^{M,g} = t$ iff for some $a \in D_A\big(\|\|\phi\|\|^{M,g[u/a]} = t\big)$<br>• $\phi_\Pi$ is true in $M$ (false in $M$) iff $\|\|\phi\|\|^{M,g} = t$ ($f$) for all $g$.<br>• $\phi_\Pi$ is logically true (false) iff $\phi$ is true (false) in every $M$<br>• $\phi_\Pi \models \psi_\Pi$ iff for every $M$ such that $\phi$ is true in $M$, $\psi$ is true in $M$ |

Table 2: Interpretation of IHTT expressions

correspond to logic formulas. A word or phrase $x$ is linked to formulas for sequences $uxv$ in which it occurs, and each formula $F$ is generalized to other formulas $G$ that entail $F$. But it is not clear yet how this representation could be used for inferences.

**Distributions, extensions, and intensions** Like the current paper, Copestake and Herbelot (2012) consider the connection between distributional representations and the semantics of logical languages. However, they reach a very different conclusion. They propose using distributional representations as intensions of logical expressions. In addition, they link distributions to extensions by noting that each sentence that contributes to the distributional representation for the word "woodchuck" is about some member of the extension of $woodchuck$. They define the *ideal distribution* for a concept, for example "woodchuck", as the collection of all true statements about all members of the category, in this case all woodchucks in the world.

In our view, distributions describe general, intensional knowledge, and do not provide reference to extensions, so we will link distributional representations to intensions and not extensions. Concerning the Copestake and Herbelot proposal of distributions as intensions, we consider distributions as representations in need of an interpretation or intension, rather than representations that constitute the intension. [3] Also it is a somewhat unclear how the intension would be defined in practice in the Copestake and Herbelot framework, as it is based on the hypothetical ideal distribution with its potentially infinite number of sentences.

**Hyper-intensional semantics** The axiom of Extensionality states that if two expressions have the same extension, then they share all their properties. Together with the standard formulation of intensions as functions from possible worlds to extensions, this generates the problem that logically equivalent statements like "John sleeps" and "John sleeps, and Mary runs or does not run" become intersubstitutable in

[3] Though it should be noted that there is a debate within psychology on whether mental conceptual knowledge is actually distributional in nature (Landauer and Dumais, 1997; Barsalou, 2008; Andrews et al., 2009).

all contexts, even in contexts like "Sue believes that..." where they should not be exchangeable. Hyperintensional semantics addresses this problem. In particular, some approaches (Fox and Lappin, 2001, 2005; Muskens, 2007) address the problem by (1) dropping the axiom of Extensionality, (2) mapping expressions of the logic first to intensions and then mapping the intensions to extensions, and (3) adopting a notion of intensions as abstract objects with minimal restrictions. This makes these approaches relevant for our purposes, as we can add the axioms that we need for a joint semantics of logical and distributional representations. Muskens (2007) has one constraint on intensions that makes the approach unusable for our purposes in its current form: It has intensions and extensions be objects from the same collections of domains – but we would not want to force extensions to be mental objects. Instead we build on the intensional higher-order type theory IHTT from Fox and Lappin (2001). The set of types of IHTT contains the basic types $e$ (for entity) and $\Pi$ (proposition), and if $A$, $B$ are types, then so is $\langle A, B\rangle$. The logic contains all the usual connectives, plus "$\cong$" for extensional equality and "=" for intensional equality. Fox and Lappin adopt the axioms shown in Table 1, which do not include the axiom of Extensionality. [4] A model for IHTT is a tuple $M = \langle D, S, L, I, F\rangle$, where $D$ is a family of non-empty sets such that $D_A$ is the set of possible extensions for expressions of type $A$. $S$ is the set of possible intensions, and $L \subseteq S$ is the set of possible intensions for non-logical constants of the logic. $I$ is a function that maps arbitrary expressions of IHTT to the set $S$ of intensions. If $\alpha$ is a non-logical constant, then $I(\alpha)$ is in $L$, otherwise $I(\alpha)$ is in $S - L$. The function $F$ is a mapping from $L$ (intensions of non-logical constants) to members of $D$ (extensions). A valuation $g$ is a function from the variables of IHTT to members of $D$ such that for all $v_A$ it holds that $g(v) \in D_A$. A model of IHTT has to satisfy the following constraints: [5]

(M1) If $v$ is a variable, then $I(v) = v$.

(M2) For a model $M$, if $I(\alpha) = I(\beta)$, then for all $g$, $||\alpha||^{M,g} = ||\beta||^{M,g}$.

Table 2 shows the definition of extensions $||.||^{M,g}$ of expressions of IHTT.

# 3 A joint semantics for distributional and logical representations

In this section we construct a first implementation of the semantics for distributional representations sketched in the introduction. In this semantics, distributional interpretations are interpreted over mental concepts and are linked to the intensions of some logical expressions. We use as a basis the hyperintensional logic IHTT of Fox and Lappin (2001) (Section 2), which does not require intensions to be mappings from possible worlds to extensions, such that we are free to link intensions to mental concepts. The central result of this section will be that the interpretation of sentences of the logic is invariant to rewriting steps such as the one in Figure 1, which replace a non-logical constant by another based on distributional similarity. The semantics that we present in this paper constitutes a first step. It leaves some important questions open, such as paraphrasing beyond the word level, or graded concept membership.

## 3.1 Distributional representations

Typically, the distributional representation for a target word $t$ is computed from the occurrences, or *usages*, of $t$ in a given corpus. Minimally, a usage is a sequence of words in which the target appears at least once. We will allow for two additional pieces of information in a usage, namely larger discourse context, and non-linguistic context. (Recently, there have been distributional approaches that make use of non-linguistic context, in particular image data (Feng and Lapata, 2010; Bruni et al., 2012).)

Let $W$ be a set of words (the lexicon), and let $Seq(W)$ be the set of finite sequences over $W$. Then a *usage* over $W$ is a tuple $\langle s, t, \delta, \omega\rangle$, where $s \in Seq(W)$ is a sequence of words such that a word form of $t \in W$ occurs in $s$ at least once, $\delta \in \Delta \cup \{NA\}$ is a (possibly empty) discourse context, and

[4] We write $\alpha_A$ to indicate that expression $\alpha$ is of type $A$.

[5] Fox and Lappin mention that one could add the constraint that if $\alpha, \alpha'$ differ only in the names of bound variables, then $I(\alpha) = I(\alpha')$. We do not do that here, since we are only concerned with replacing non-logical constants in the current paper.

$\omega \in \Omega \cup \{NA\}$ is a (possibly empty) non-linguistic context. We write $\mathcal{U}(W, \Delta, \Omega)$ for the set of all usages over $W$ (and $\Delta$ and $\Omega$). For any usage $u = \langle s, t, \delta, \omega \rangle$, we write $target(u) = t$. Given a set $U \subseteq \mathcal{U}(W, \Delta, \Omega)$ of usages, we write $U_t = \{u \in U \mid target(u) = t\}$ for the usages of a target word $t$. Furthermore, we write $W_U = \{t \in W \mid U_t \neq \emptyset\}$ for the set of words that have usages in $U$.

In distributional approaches, the vector space representation for a target word $t$ is computed from such a set $U$ of usages, typically by mapping $U$ to a single point in vector space (Lund et al., 1995; Landauer and Dumais, 1997) or a set of points (Schütze, 1998; Reisinger and Mooney, 2010). This makes it possible to use linear algebra in modeling semantics. However, for our current purposes, we do not need to specify any particular mapping to a vector space, and can simply work with the underlying set $U$ of usages: A finite set $U$ of usages over $W$ constitutes a *distributional representation* for $W_U$. The distributional representation for a word $t \in W$ is $U_t$.

### 3.2 A semantics for distributional representations

We want to interpret distributional representations over conceptual structure. But what is conceptual structure? We know that concepts are linked by different semantic relations, including is-a, and part-of (Fellbaum, 1998), they can overlap, and they are associated with definitional features (Murphy, 2002). Eventually, all of these properties may be useful to include in the semantics of distributional representations. But for this first step we work with a much simpler definition. We define a conceptual structure simply as a set of (atomic, unconnected) concepts.

An individual usage of a word $t$ can refer to a single mental concept. For example, the usage of "bank" in (1) clearly refers to a "financial institution" concept, not the land at the side of a river. But an individual usage can also refer to multiple mental concepts when there is ambiguity as in (2), or when there is too little information to determine the intended meaning as in (3). [6][7]

(1) $\langle$The bank engaged in risky stock trades, bank, $\delta$, $\omega\rangle$

(2) $\langle$Why fix dinner when it isn't broken, fix, $\delta$, $\omega\rangle$[8]

(3) $\langle$bank, bank, NA, NA$\rangle$

From this link between individual usages and concepts, we can derive a link between distributional representations and concepts: The representation $U_t$ of a word $t$ is connected to all concepts to which the usages in $U_t$ link. Formally, a *conceptual model* for $\mathcal{U}(W, \Delta, \Omega)$ is a tuple $\mathcal{C} = \langle I_u, C \rangle$, where $C$ is a set of concepts, and the function $I_u : \mathcal{U}(W, \Delta, \Omega) \rightarrow 2^C$ is an interpretation function for usages that maps each usage to a set of concepts. [9] A conceptual model $\mathcal{C}$ together with a finite set $U \subseteq \mathcal{U}(W, \Delta, \Omega)$ of usages define a conceptual mapping for words. We write $I_{\mathcal{C},U}(w) = \bigcup_{u \in U_w} I_u(u)$ for the set of concepts associated with $w$.

Distributional approaches centrally use some similarity measure, for example cosine similarity, on pairs of distributional representations, usually pairs of points in vector space. Since we represent a word $t$ directly by its set $U_t$ of usages rather than a point in vector space derived from $U_t$, we instead have a similarity measure $sim(U_1, U_2)$ on sets of usages. We assume a range of $[0, 1]$ for this similarity measure. A conceptual model can be used to evaluate the appropriateness of similarity predictions: A prediction is appropriate if it is high for two usage sets that refer to the same concepts, or low for two usage sets that refer to different concepts. Formally, a similarity prediction $sim(U_1, U_2)$ is *appropriate* for a conceptual model $\mathcal{C} = \langle I_u, C \rangle$ and threshold $\theta$ iff

- either $sim(U_1, U_2) \geq \theta$ and $\bigcup_{u \in U_1} I_u(u) = \bigcup_{u \in U_2} I_u(u)$,

[6] For the purpose of this paper we make the simplifying assumption that concepts have "strict boundaries": A usage either does or does not refer to a concept. We do not model cases where a usage is related to a concept, but is not a clear match.

[7] Another possible reason for one usage mapping to multiple mental concepts is concept overlap (Murphy, 2002).

[8] Advertisement for a supermarket in Austin, Texas..

[9] We write $2^S$ for the power set of a set $S$.

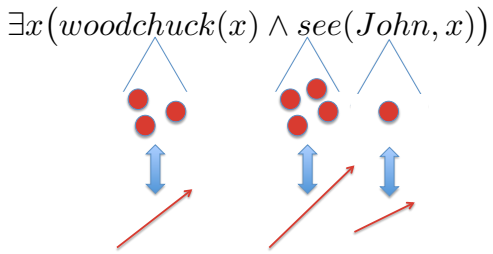


Figure 2: Enriching the information about non-logical constants: Constants are associated with sets of concepts (circles) and, through them, with distributional representations

- or $sim(U_1, U_2) < \theta$ and $\bigcup_{u \in U_1} I_u(u) \neq \bigcup_{u \in U_2} I_u(u)$.

This formulation of appropriateness is simplistic in that it discretizes similarity predictions into two classes: above or below threshold $\theta$. This is due to our current impoverished view of concepts as disjoint atoms. When we introduce a conceptual similarity measure within conceptual models, a more fine-grained evaluation of distributional similarity ratings becomes available. Such a conceptual similarity measure would be justified, as humans can judge similarity between concepts (Rubenstein and Goodenough, 1965), but we do not do it here in order to keep our models maximally simple.

### 3.3 A joint semantics for logical form and distributional representations

We now link the intensions of some logical expressions to mental concepts, using the logic IHTT as a basis. We will need to constrain the behavior of intensions more than Fox and Lappin do. In particular, we add the following two requirements to models $M = \langle D, S, L, I, F\rangle$ of IHTT.

**(M3)** If the expression $\alpha \in A$ is the result of beta-reducing the expression $\beta \in A$, then $I(\alpha) = I(\beta)$.

**(M4)** If $I(u_A) = I(v_A)$, then for all $\phi \in \langle A, B\rangle$, $I(\phi(u)) = I(\phi(v))$.

(M4) allows for the exchange of an intensionally equal expression without changing the intension of the overall expression.

We now define models that join an intensional model of IHTT with a conceptual model for a distributional representation. In particular, we link constants of the logic to sets of concepts, and through them, to distributional representations, as sketched in Figure 2. If the word "woodchuck" is associated with the concept set $C_{\text{woodchuck}} = I_{\mathcal{C},U}(\text{woodchuck})$, then the intension of the constant $woodchuck$ will also be $C_{\text{woodchuck}}$. We proceed in two steps: In the definition of joint models, we require the existence of a mapping from words to non-logical constants that share the same interpretation. In a second step, we require semantic constructions to respect this mapping, such that the logical expression associated with "woodchuck" will be $\lambda x\big(woodchuck(x)\big)$ rather than $\lambda x\big(guppy(x)\big)$. Note that only words in $W_U$ have distributional representations associated with them; for words in $W - W_U$, neither their translation to logical expressions nor the intensions of those expressions are constrained in any way.

Let $M = \langle D, S, L, I, F\rangle$ be a model for IHTT, let $\mathcal{C} = \langle I_u, C\rangle$ be a conceptual model for $\mathcal{U}(W, \Delta, \Omega)$, and let $U$ be a finite subset of $\mathcal{U}(W, \Delta, \Omega)$. Then $M_{\mathcal{C}} = \langle D, S, L, I, F, I_u, C\rangle$ is an *intensional conceptual model* for IHTT and $\mathcal{U}(W, \Delta, \Omega)$ *based on* $U$ if

**(M5)** There exists a function $h$ from $W_U$ to the non-logical constants of IHTT such that for all $w \in W_U$, $I_{\mathcal{C},U}(w) = I(h(w))$

**(M6)** For all $w_1, w_2 \in W_U$, if $I_{\mathcal{C},U}(w_1) = I_{\mathcal{C},U}(w_2)$ then $h(w_1)$ and $h(w_2)$ have the same type.

We say that the model $M_{\mathcal{C}}$ *contains* $M$ and $\mathcal{C}$.

Constraint (M5) links each word to a non-logical constant such that the distributional interpretation of the word and the intension of the constant are the same. (M6) states that if two words have the same

distributional interpretation, their associated constants have the same type. We next define semantic constructions $sem$ in general, and semantic constructions that connect the translation $sem(w)$ of a word $w$ to its associated constant $h(w)$. A *semantic construction function* for a set $W$ of words and a logical language $\mathcal{L}$ is a partial function $sem : Seq(W) \rightarrow \mathcal{L}$ such that $sem(w)$ is defined for all $w \in W$. $sem(.)$ maps sequences of words over $W$ to expressions from $\mathcal{L}$. A sequence $s \in Seq(W)$ is called *grammatical* if $sem(s)$ is defined. A semantic construction $sem$ is an *intended semantic construction* for an intensional conceptual model $M = \langle D, S, L, I, F, I_u, C\rangle$ based on $U$ if the following constraint holds for the function $h$ from (M5):

**(M7)** For each type $A$ there exists some expression $\phi^A$ such that for all $w \in W_U$, $sem(w)$ is equivalent (modulo beta-reduction) to $\phi^A(h(w))$.

(M7) states that the construction of translations $sem(w)$ from non-logical constants $h(w)$ must be uniform for all words of the same semantic type. For example, if for the word "woodchuck" we have $h(\text{woodchuck}) = woodchuck$, an expression of type $\langle e, \Pi\rangle$, then the expression $\phi_{\langle e,\Pi\rangle} = \lambda P \lambda x(P(x))$ will map $woodchuck$ to $\lambda x(woodchuck(x)) = sem(\text{woodchuck})$.

### 3.4 Synonym replacement

In Section 2 we have sketched a framework for the interaction of logic and distributional representations based on Bar-Haim et al. (2007). Distributional representations can be used to predict semantic similarity between pairs of words and in particular to predict synonymy between words (Lin, 1998). Distributionally induced synonym pairs can be used as rewriting rules that transform sentence representations. In our case, the representations to be transformed are expressions of the logic. Two sentences count as synonymous if it is possible to transform the representation of one sentence into the representation of the other, using both distributional rewriting rules and the axioms of the logic.

We start out by showing that the application of a rewriting rule that exchanges one non-logical constant of IHTT for another constant with the same intension leaves both the intension and the extension of the overall logical expression unchanged. Given a logical expression $\phi$, we write $\phi[\text{some } b/a]$ for the set of expressions obtained from $\phi$ by replacing zero or more occurrences of $a$ by $b$.

**Proposition 1: Soundness of non-logical constant rewriting.** Let $M = \langle D, S, L, I, F\rangle$ be an intensional model for IHTT, and let $a, b$ be non-logical constants of IHTT of type $A$ such that $I(a) = I(b)$. Then for any expression $\phi$ of IHTT and any $\phi' \in \phi[\text{some } b/a]$, $I(\phi) = I(\phi')$, and for any valuation $g$, $||\phi||^{M,g} = ||\phi'||^{M,g}$.

**Proof.** Let $x_A$ be a variable that does not occur in $\phi$. Then for each $\phi' \in \phi[\text{some } b/a]$ there exists an expression $\psi \in \phi[\text{some } x/a]$ such that $(\lambda x\psi)(a)$ beta-reduces to $\phi$ and $(\lambda x\psi)(b)$ beta-reduces to $\phi'$. As $I(a) = I(b)$, we have $I((\lambda x\psi)(a)) = I((\lambda x\psi)(b))$ by (M4). So by (M3), $I(\phi) = I(\phi')$. From this it follows that for any valuation $g$, $||\phi||^{M,g} = ||\phi'||^{M,g}$ by (M2). □

We call two words synonyms if they refer to the same set of concepts. Formally, let $U$ be a finite subset of $\mathcal{U}(W, \Delta, \Omega)$ that is a distributional representation for $W_U$, and $\mathcal{C} = \langle I_u, C\rangle$ a conceptual model for $\mathcal{U}(W, \Delta, \Omega)$. A word $p \in W_U$ is a *synonym* for $t \in W_U$ by $\mathcal{C}$ and $U$ if $I_{\mathcal{C},U}(t) = I_{\mathcal{C},U}(p)$.

We would like to show that if $t$ and $p$ are synonyms, then exchanging $t$ for $p$ changes neither the intension nor the extension of the logical translation for the sentence. To do so, we first show that exchanging $t$ for $p$ corresponds to applying constant rewriting on the sentence representation.

Note, however, that the logical translation of a sentence depends not only on the words, but also on the syntactic structure of the sentence. If a given syntactic analysis framework only allows for the bracketing "(small (tree house))" and at the same time only allows for the bracketing "((little tree) house)", then the two phrases will not receive the same semantics even if the model considers "small" and "little" to be synonyms. So we will show that if replacement by a synonym *within a given syntactic structure* again yields a valid syntactic structure, then the semantics of the sentence remains unchanged. For any

sequence $s \in Seq(W)$ of words over $W$, we write $T(s)$ for the set of constituent structure analyses for $s$. For $\tau \in T(s)$, we write $\tau[p/t]$ for the syntactic graph that is exactly like $\tau$ except that all leaves labeled $t$ are replaced by leaves labeled $p$. We write $sem(\tau)$ for the logical translation of $s$ that is based on the syntactic structure of $\tau$. We assume that there exists exactly one translation $sem(\tau)$ for each syntactic structure $\tau$.

**Lemma 2.** Let $M_{\mathcal{C}}$ be be an intensional conceptual model for IHTT and $\mathcal{U}(W, \Delta, \Omega)$ based on $U \subseteq \mathcal{U}(W, \Delta, \Omega)$ that contains $M = \langle D, S, L, I, F\rangle$ and $\mathcal{C} = \langle I_u, C\rangle$. Let $t, p \in W_U$ be synonyms by $\mathcal{C}$ and $U$, and let $s \in Seq(W)$ be a sequence with syntactic analysis $\tau \in T(s)$ such that $\tau[p/t] \in T(s[p/t])$. Then for any intended semantic construction $sem$ for $M_{\mathcal{C}}$ and $U$, $sem(\tau[p/t])$ is equivalent modulo beta-reduction to some member of $sem(\tau)[\text{some } h(p)/h(t)]$.

**Proof.** We proceed by induction over the structure of $\tau$. If $s$ consists of a single word, then $\tau = s$, and either $s = t$ or $s = w$ for a word $w \neq t$. If $s = w$ for some $w \neq t$, then $sem(\tau[p/t]) = sem(\tau) \in sem(\tau)[\text{some } h(p)/h(t)]$.

If $s = t$, then $sem(\tau) = sem(t)$ and $sem(\tau[p/t]) = sem(p)$. By (M5) and because $t$ and $p$ are synonyms, we have $I(h(t)) = I_{\mathcal{C},U}(t) = I_{\mathcal{C},U}(p) = I(h(p))$. From this it follows by (M6) that the non-logical constants $h(t)$ and $h(p)$ have the same semantic type $A$. Then by (M7) there exists a logical expression $\phi^A$ such that $sem(\tau) = sem(t)$ is equivalent modulo beta-reduction to $\phi^A(h(t))$. At the same time, $sem(\tau[p/t]) = sem(p)$ is equivalent modulo beta-reduction to $\phi^A(h(p))$, which is equivalent modulo beta-reduction to a member of $\big(\phi^A(h(t))\big)[\text{some } h(p)/h(t)]$, which in turn is equivalent modulo beta-reduction so a member of $sem(\tau)[\text{some } h(p)/h(t)]$.

Now assume that $s$ comprises more than one word. Let the root of $\tau$ have $n$ children that are the roots of subtrees $\tau_1 \ldots \tau_n$. There is some semantic construction rule associated with the root of $\tau$ that can be written as an expression $\phi$ of IHTT such that $\phi(sem(\tau_1)) \ldots (sem(\tau_n))$ beta-reduces to $sem(\tau)$. By the inductive hypothesis, $sem(\tau_i[p/t])$ is equivalent modulo beta-reduction to some $\psi_i \in sem(\tau_i)[\text{some } h(p)/h(t)]$ for $1 \leq i \leq n$. The expression $\phi$ remains unchanged between $sem(\tau)$ and $sem(\tau[p/t])$ because only leaves of the tree were changed and the overall constituent structure remained the same. So the expression $sem(\tau[p/t])$ is equivalent modulo beta-reduction to $\phi(\psi_1) \ldots (\psi_n) \in \big(\phi(sem(\tau_1)) \ldots (sem(\tau_n))\big)[\text{some } h(p)/h(t)]$, which in turn is equivalent modulo beta-reduction to $sem(\tau)[\text{some } h(p)/h(t)]$. □

The reason why we have used $\phi[\text{some } b/a]$ rather than replacement of all occurrences is that there is no guarantee that the corresponding non-logical constant $h(t)$ for a word $t$ is used only in the lexical entry of $t$. For example, the expression $\phi^{\langle e,\Pi\rangle}$ of (M7) could be $\lambda P \lambda x\big(woodchuck(x) \wedge P(x)\big)$, making the lexical entry for "guppy" $\lambda x\big(woodchuck(x) \wedge guppy(x)\big)$. Or the semantic construction expression $\phi$ for NPs could contain the constant $woodchuck$. However, now we are in a position to show that this does not matter, and that a constant rewriting rule can be applied to all occurrences of $h(t)$, whether in the lexical entry for $t$ or elsewhere. At the same time, we show that replacement of a word by a synonym does not change the interpretation of the sentence.

**Proposition 3: Synonym replacement as constant replacement.** Let $M_{\mathcal{C}}$ be be an intensional conceptual model for IHTT and $\mathcal{U}(W, \Delta, \Omega)$ based on $U \subseteq \mathcal{U}(W, \Delta, \Omega)$ that contains $M = \langle D, S, L, I, F\rangle$ and $\mathcal{C} = \langle I_u, C\rangle$. Let $t, p \in W_U$ be synonyms by $\mathcal{C}$ and $U$, and let $s \in Seq(W)$ be a sequence with syntactic analysis $\tau \in T(s)$ such that $\tau[p/t] \in T(s[p/t])$. Then for any valuation $g$, and any intended semantic construction $sem$ for $M_{\mathcal{C}}$ and $U$, $I(sem(\tau)) = I(sem(\tau[p/t])) = I(sem(\tau)[h(p)/h(t)])$, and $||sem(\tau)||^{M,g} = ||sem(\tau[p/t])||^{M,g} = ||sem(\tau)[h(p)/h(t)]||^{M,g}$.

**Proof.** By Lemma 2, the semantic representation of the changed syntactic tree, $sem(\tau[p/t])$, is equivalent modulo beta-reduction to some $\psi \in sem(\tau)[\text{some } h(p)/h(t)]$. So by Proposition 1, $I(\psi) =$

$I(sem(\tau))$, and by (M3), $I(sem(\tau)[p/t]) = I(\psi)$. Thus, $I(sem(\tau)) = I(sem(\tau[p/t]))$. By Proposition 1, the intension is the same for all members of $sem(\tau)$[some $h(p)/h(t)$], so we have $I(sem(\tau)) = I(sem(\tau)[h(p)/h(t)]$. And by (M2), if $sem(\tau)$, $sem(\tau[p/t])$ and $sem(\tau)[h(p)/h(t)]$ have the same intension, they also have the same extension. □

### 3.5 Inference

We extend the list of axioms for IHTT from Table 1 by two additional axioms that correspond to the constraints (M3) and (M4).

**(IHTT14)** $\vdash \lambda u \phi(v) = \phi[u/v]$ (where $u$ is a variable in $A$, $v \in A$, $\phi \in \langle A, B \rangle$, and $v$ is not bound when substituted for $u$ in $\phi$)

**(IHTT15)** $\vdash \forall u, v_A \forall \phi_{\langle A,B \rangle} \big(u = v \rightarrow \phi(u) = \phi(v)\big)$

These axioms parallel (IHTT9) and (IHTT12) but state intensional rather than extensional equality.

Synonymy predictions from the distributional representation can be transformed into rewriting rules: If the words $t$ and $p$ are synonyms by the distributional representation $U$, then we generate the rewriting rule $h(t) \mapsto h(p)$. As Proposition 3 shows, this rewriting rule can be applied indiscriminately to a logical expression, and is not restricted to the lexical entry for $t$. But since the logic is equipped with inference capability and is not a passive representation like the syntactic graphs that Bar-Haim et al. (2007) used, we can alternatively just inject an expression $h(t) = h(p)$, which states intensional equality, into the logical representation for the parsed sentence $\tau$. The logical representation for $\tau[p/t]$ can then be inferred using (IHTT14) and (IHTT15).

## 4 Conclusion and outlook

In this paper we have proposed a semantics for distributional representations, namely that each point in vector space stands for a set of mental concepts. We have provided a coarse-grained evaluation for distributional representations in which their similarity predictions are evaluated against conceptual equality or inequality. We have extended this approach to a joint semantics of distributional and logical representations by linking the intensions of some logical expressions to mental concepts as well: If the distributional representation for a word $w$ is interpreted as a set $C$ of concepts, then the non-logical constant linked to the lexical entry for $w$ will have as its intension the same set $C$. We have used hyperintensional semantics as a basis for this joint semantics. We have been able to show that distributional rewriting rules that exchange non-logical constants with the same intension do not change the intension or extension of the overall logical expression. These rewriting rules can be used to compute the logical representation of a sentence after exchanging a word for its synonym.

The current joint semantics is, however, only a first step, and leaves many important questions open. We consider the following three to be especially important. *(1) Polysemy.* Many synonym pairs can only be substituted for one another in particular sentence contexts. For example "correct" is a synonym for "fix" that can be substituted in the context of "The programmer fixed the error", but not in "The cook fixed dinner." This means that the words "fix" and "correct" do not map to the same set of concepts, but they are exchangeable in particular contexts. So we would want to say that "fix" and "correct" are synonyms with respect to a usage $u = \langle s, \text{fix}, \delta, \omega \rangle$ if $I_u(u) = I_u(\langle s[\text{correct}/\text{fix}], \text{correct}, \delta, \omega \rangle)$. The main challenge for incorporating polysemy is to have intensions change based on the context of use.

*(2) Distributional similarity of larger phrases.* There is considerable work both on the distributional similarity of phrases and sentences (Coecke et al., 2010; Baroni and Zamparelli, 2010; Grefenstette and Sadrzadeh, 2011) and on the distributional similarity of phrases with open argument slots, such as "X solves Y" and "X finds a solution to Y" (Lin and Pantel, 2001; Szpektor and Dagan, 2008; Berant et al., 2011). We would like to use these results to do distributionally driven replacement of multi-word paraphrases in a joint distributional and logical framework. But this requires a semantics for distributional representations of larger phrases. If we assume some sort of conceptual structures as semantics, the next

question is whether all logical expressions should be associated with conceptual structures: Should the intension of a variable be something conceptual?

*(3) Gradience.* In this paper we have assumed that the link from usage to concept is binary – either present or not –, and also that there are no relations between concepts. Both assumptions are simplifications: Concepts have "fuzzy boundaries" (Hampton, 2007), and cognizers can distinguish degrees of similarity between concepts (Rubenstein and Goodenough, 1965). By modeling this gradience, we could then talk about degrees of similarity between words and phrases, not just a binary choice of either synonymy or non-synonymy. But this will require dealing with probabilities or weights in the model and also in the logic.

**Acknowledgements.** This research was supported in part by the NSF CAREER grant IIS 0845925 and by the DARPA DEFT program under AFRL grant FA8750-13-2-0026. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the view of DARPA, AFRL or the US government. Warmest thanks to John Beavers and Gemma Boleda, as well as the members of the Austin Computational Linguistics Tea and the anonymous reviewers, for very helpful discussions.

## References

Andrews, M., G. Vigliocco, and D. Vinson (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review 116*(3), 463–498.

Bar-Haim, R., I. Dagan, I. Greental, and E. Shnarch (2007). Semantic inference at the lexical-syntactic level. In *Proceedings of AAAI*, Vancouver, Canada.

Baroni, M. and R. Zamparelli (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, Cambridge, MA.

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology 59*(1), 617–645.

Berant, J., I. Dagan, and J. Goldberger (2011). Global learning of typed entailment rules. In *Proceedings of ACL*, Portland, OR.

Blei, D. M., A. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Bruni, E., G. Boleda, M. Baroni, and N. Tran (2012). Distributional semantics in technicolor. In *Proceedings of ACL*, Jeju Island, Korea.

Burgess, C. and K. Lund (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes 12*, 177–210.

Clark, S., B. Coecke, and M. Sadrzadeh (2008). A compositional distributional model of meaning. In *Proceedings of QI*, Oxford, UK, pp. 133–140.

Clarke, D. (2007). *Context-theoretic Semantics for Natural Language: an Algebraic Framework*. Ph. D. thesis, University of Sussex.

Clarke, D. (2012). A context-theoretic framework for compositionality in distributional semantics. *Computational Linguistics 38*(1).

Coecke, B., M. Sadrzadeh, and S. Clark (2010). Mathematical foundations for a compositional distributed model of meaning. *Lambek Festschrift, Linguistic Analysis 36*.

Copestake, A. and A. Herbelot (2012, July). Lexicalised compositionality. Unpublished draft.

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

Feng, Y. and M. Lapata (2010). Visual information in semantic representation. In *Proceedings of HLT-NAACL*, Los Angeles, California.

Fox, C. and S. Lappin (2001). A framework for the hyperintensional semantics of natural language with two implementations. In P. de Groote, G. Morrill, and C. Retore (Eds.), *Proceedings of LACL*, Le Croisic, France.

Fox, C. and S. Lappin (2005). *Foundations of Intensional Semantics*. Wiley-Blackwell.

Gärdenfors, P. (2004). *Conceptual spaces*. Cambridge, MA: MIT press.

Garrette, D., K. Erk, and R. Mooney (2011). Integrating logical representations with probabilistic information using markov logic. In *Proceedings of IWCS*, Oxford, UK.

Grefenstette, E. and M. Sadrzadeh (2011). Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of EMNLP*, Edinburgh, Scotland, UK.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science 31*, 355–384.

Landauer, T. and S. Dumais (1997). A solution to Platos problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*(2), 211–240.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*, Montreal, Canada.

Lin, D. and P. Pantel (2001). Discovery of inference rules for question answering. *Natural Language Engineering 7*(4), 343–360.

Lund, K., C. Burgess, and R. Atchley (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the Cognitive Science Society*, pp. 660–665.

Murphy, G. L. (2002). *The Big Book of Concepts*. MIT Press.

Muskens, R. (2007). Intensional Models for the Theory of Types. *The Journal of Symbolic Logic 72*(1), 98–118.

Reisinger, J. and R. Mooney (2010). Multi-prototype vector-space models of word meaning. In *Proceeding of NAACL*.

Rubenstein, H. and J. Goodenough (1965). Contextual correlates of synonymy. *Computational Linguistics 8*, 627–633.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics 24*(1).

Socher, R., E. Huang, J. Pennin, A. Ng, and C. Manning (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (Eds.), *Proceedings of NIPS*.

Szpektor, I. and I. Dagan (2008). Learning entailment rules for unary templates. In *Proceedings of COLING*.

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

Turney, P. D. (2012). Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research 44*, 533–585.