A Quantum-Theoretic Approach to Distributional Semantics

William Blacoe[1], **Elham Kashefi**[2], **Mirella Lapata**[1]
[1]Institute for Language, Cognition and Computation,
[2]Laboratory for Foundations of Computer Science
School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB
w.b.blacoe@sms.ed.ac.uk, ekashefi@inf.ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

In this paper we explore the potential of quantum theory as a formal framework for capturing lexical meaning. We present a novel semantic space model that is syntactically aware, takes word order into account, and features key quantum aspects such as superposition and entanglement. We define a dependency-based Hilbert space and show how to represent the meaning of words by density matrices that encode dependency neighborhoods. Experiments on word similarity and association reveal that our model achieves results competitive with a variety of classical models.

1 Introduction

The fields of cognitive science and natural language processing have recently produced an ensemble of semantic models which have an impressive track record of replicating human behavior and enabling real-world applications. Examples include simulations of word association (Denhière and Lemaire, 2004; Griffiths et al., 2007), semantic priming (Lund and Burgess, 1996; Landauer and Dumais, 1997; Griffiths et al., 2007), categorization (Laham, 2000), numerous studies of lexicon acquisition (Grefenstette, 1994; Lin, 1998), word sense discrimination (Schütze, 1998), and paraphrase recognition (Socher et al., 2011). The term "semantic" derives from the intuition that words seen in the context of a given word contribute to its meaning (Firth, 1957). Although the specific details of the individual models differ, they all process a corpus of text as input and represent words (or concepts) in a (reduced) high-dimensional space.

In this paper, we explore the potential of quantum theory as a formal framework for capturing lexical meaning and modeling semantic processes such as word similarity and association (see Section 6 for an overview of related research in this area). We use the term *quantum theory* to refer to the abstract mathematical foundation of quantum mechanics which is not specifically tied to physics (Hughes, 1989; Isham, 1989). Quantum theory is in principle applicable in any discipline where there is a need to formalize uncertainty. Indeed, researchers have been pursuing applications in areas as diverse as economics (Baaquie, 2004), information theory (Nielsen and Chuang, 2010), psychology (Khrennikov, 2010; Pothos and Busemeyer, 2012), and cognitive science (Busemeyer and Bruza, 2012; Aerts, 2009; Bruza et al., 2008). But what are the features of quantum theory which make it a promising framework for modeling meaning?

Superposition, entanglement, incompatibility, and interference are all related aspects of quantum theory, which endow it with a unique character.[1] *Superposition* is a way of modeling uncertainty, more so than in classical probability theory. It contains information about the potentialities of a system's state. An electron whose location in an atom is uncertain can be modeled as being in a superposition of locations. Analogously, words in natural language can have multiple meanings. In isolation, the word *pen* may refer to a writing implement, an enclosure for confining livestock, a playpen, a penitentiary or a female swan. However, when observed in the context of the word *ink* the ambiguity resolves into the sense of the word dealing with writing. The meanings of words in a semantic space are superposed in a way which is intuitively similar to the atom's electron.

Entanglement concerns the relationship between

[1]It is outside the scope of the current paper to give a detailed introduction on the history of quantum mechanics. We refer the interested reader to Vedral (2006) and Kleppner and Jackiw (2000) for comprehensive overviews.

Proceedings of NAACL-HLT 2013, pages 847–857,
Atlanta, Georgia, 9–14 June 2013. ©2013 Association for Computational Linguistics

systems for which it is impossible to specify a joint probability distribution from the probability distributions of their constituent parts. With regard to word meanings, entanglement encodes (hidden) relationships between concepts. The different senses of a word "exist in parallel" until it is observed in some context. This reduction of ambiguity has effects on other concepts connected via entanglement. The notion of *incompatibility* is fundamental to quantum systems. In classical systems, it is assumed by default that measurements are compatible, that is, independent, and as a result the order in which these take place does not matter. By contrast in quantum theory, measurements may share (hidden) order-sensitive inter-dependencies and the outcome of the first measurement can change the outcome of the second measurement.

Interference is a feature of quantum probability that can cause classical assumptions such as the law of total probability to be violated. When concepts interact their joint representation can exhibit non-classical behavior, e.g., with regard to conjunction and disjunction (Aerts, 2009). An often cited example is the "guppy effect". Although *guppy* is an example of a *pet-fish* it is neither a very typical *pet* nor *fish* (Osherson and Smith, 1981).

In the following we use the rich mathematical framework of quantum theory to model semantic information. Specifically, we show how word meanings can be expressed as quantum states. A word brings with it its own subspace which is spanned by vectors representing its potential usages. We present a specific implementation of a semantic space that is syntactically aware, takes word order into account, and features key aspects of quantum theory. We empirically evaluate our model on word similarity and association and show that it achieves results competitive with a variety of classical models. We begin by introducing some of the mathematical background needed for describing our approach (Section 2). Next, we present our semantic space model (Section 3) and our evaluation experiments (Sections 4 and 5). We conclude by discussing related work (Section 6).

2 Preliminaries

Let $c = r e^{i\theta}$ be a complex number, expressed in polar form, with absolute value $r = |c|$ and phase θ. Its complex conjugate $c^* = r e^{-i\theta}$ has the inverse phase. Thus, their product $cc^* = (r e^{i\theta})(r e^{-i\theta}) = r^2$ is real.

2.1 Vectors

We are interested in finite-dimensional, complex-valued vector spaces $\mathbb{C}^n$ with an inner product, otherwise known as Hilbert space. A column vector $\overrightarrow{\psi} \in \mathbb{C}^n$ can be written as an ordered vertical array of its n complex-valued components, or alternatively as a weighted sum of base vectors:

$$\overrightarrow{\psi} = \begin{pmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \end{pmatrix} = \psi_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \ldots + \psi_n \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \quad (1)$$

Whereas Equation (1) uses base vectors from the standard base $\mathcal{B}_{std} = \{\overrightarrow{b_1}, ..., \overrightarrow{b_n}\}$, any other set of n orthonormal vectors serves just as well as a base for the same space. Dirac (1939) introduced the so-called bra-ket notation which is equally expressive but notationally more convenient. A column vector becomes a ket:

$$\overrightarrow{\psi} \equiv |\psi\rangle = \psi_1|b_1\rangle + \psi_2|b_2\rangle + \ldots + \psi_n|b_n\rangle \quad (2)$$

and a row vector becomes a bra $\langle\psi|$. Transposing a complex-valued vector or matrix (via the superscript "†") involves complex-conjugating all components:

$$|\psi\rangle^\dagger = \langle\psi| = \psi_1^*\langle b_1| + \psi_2^*\langle b_2| + \ldots + \psi_n^*\langle b_n| \quad (3)$$

The Dirac notation for the inner product $\langle\cdot|\cdot\rangle$ illustrates the origin of the terminology "bra-ket". Since $\mathcal{B}_{std}$'s elements are normalised and pairwise orthogonal their inner product is:

$$\langle b_i|b_j\rangle = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

The inner product is also applicable to pairs of non-base kets:

$$\begin{aligned} &(\psi_1^* \; \psi_2^* \; \cdots \; \psi_n^*) \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_n \end{pmatrix} \equiv \langle\psi|\phi\rangle \\ &= \left(\textstyle\sum_i \psi_i^*\langle b_i|\right)\left(\textstyle\sum_j \phi_j|b_j\rangle\right) \\ &= \textstyle\sum_{i,j} \psi_i^*\phi_j\langle b_i|b_j\rangle = \sum_i \psi_i^*\phi_i\langle b_i|b_i\rangle \\ &= \textstyle\sum_i \psi_i^*\phi_i \end{aligned} \quad (5)$$

Reversing the order of an inner product complex-conjugates it:

$$(\langle\psi|\phi\rangle)^* = \langle\phi|\psi\rangle \tag{6}$$

2.2 Matrices

Matrices are sums of outer products $|\cdot\rangle\langle\cdot|$. For example, the matrix $(M_{i,j})_{i,j}$ can be thought of as the weighted sum of "base-matrices" $B_{i,j} \equiv |b_i\rangle\langle b_j|$, whose components are all 0 except for a 1 in the i-th row and j-th column. The outer product extends linearly to non-base kets in the following manor:

$$\begin{aligned} |\psi\rangle\langle\phi| &= \left(\sum_i \psi_i|b_i\rangle\right)\left(\sum_j \phi_j^*\langle b_j|\right) \\ &= \sum_{i,j} \psi_i\phi_j^*|b_i\rangle\langle b_j| \end{aligned} \tag{7}$$

This is analogous to the conventional multiplication:

$$\begin{pmatrix} \psi_1 \\ \vdots \\ \psi_n \end{pmatrix} (\phi_1^* \cdots \phi_n^*) = \begin{pmatrix} \psi_1\phi_1^* & \cdots & \psi_1\phi_n^* \\ \vdots & \ddots & \vdots \\ \psi_n\phi_1^* & \cdots & \psi_n\phi_n^* \end{pmatrix} \tag{8}$$

We will also make use of the tensor product. Its application to kets, bras and outer products is linear:

$$\begin{aligned} &(|a\rangle + |b\rangle) \otimes |c\rangle = |a\rangle \otimes |c\rangle + |b\rangle \otimes |c\rangle \\ &(\langle a| + \langle b|) \otimes \langle c| = \langle a| \otimes \langle c| + \langle b| \otimes \langle c| \\ &(|a\rangle\langle b| + |c\rangle\langle d|) \otimes |e\rangle\langle f| = \\ &(|a\rangle \otimes |e\rangle)(\langle b| \otimes \langle f|) + (|c\rangle \otimes |e\rangle)(\langle d| \otimes \langle f|) \end{aligned} \tag{9}$$

For convenience we omit "$\otimes$" where no confusion arises, e.g., $|a\rangle \otimes |b\rangle = |a\rangle|b\rangle$. When applied to Hilbert spaces, the tensor product creates the composed Hilbert space $\mathcal{H} = \mathcal{H}_1 \otimes ... \otimes \mathcal{H}_n$ whose base kets are simply induced by the tensor product of its subspaces' base kets:

$$\begin{aligned} &base(\mathcal{H}_1 \otimes ... \otimes \mathcal{H}_n) = \\ &\left\{ \bigotimes_{i=1}^{n} |b\rangle_i \ : \ |b\rangle_i \in base(\mathcal{H}_i),\ 1 \leq i \leq n \right\} \end{aligned} \tag{10}$$

Whereas the order of composed kets $|a\rangle|b\rangle|c\rangle$ usually suffices to identify which subket lives in which subspace, we make this explicit by giving subkets the same subscript as the corresponding subspace. Thus, the order no longer matters, as in the following inner product of composed kets:

$$(\langle a|_1\langle b|_2\langle c|_3)(|e\rangle_3|d\rangle_1|f\rangle_2) = \langle a|d\rangle\langle b|f\rangle\langle c|e\rangle \tag{11}$$

Definition 1. *Self-adjoint Matrix*

A matrix M is self-adjoint iff $M_{i,j} = M_{j,i}^*$ for all i, j. Consequently, all diagonal elements are real-valued, and $M = M^\dagger$ is its own transpose conjugate.

Definition 2. *Density Matrix*

A self-adjoint matrix M is a density matrix iff it is positive semi-definite, i.e., $\langle\phi|M|\phi\rangle \geq 0$ for all $|\phi\rangle \in \mathbb{C}^n$, and it has unit trace, i.e., $Tr(M) = \sum_{|b\rangle\in\mathcal{B}} \langle b|M|b\rangle = 1$.

The term "density matrix" is synonymous with "density operator". Any density matrix ρ can be decomposed arbitrarily as $\rho = \sum_i p_i|s_i\rangle\langle s_i|$, the weighted sum of sub-matrices $|s_i\rangle\langle s_i|$ with $p_i \in \mathbb{R}_{>0}$ and $\langle s_i|s_i\rangle = 1$. The p_i need not sum to 1. In fact the decomposition where the p_i sum to 1 and the $|s_i\rangle$ are mutually orthogonal is unique and is called the eigen decomposition. Consequently $\mathcal{B}_{eig} = \{|s_i\rangle\}_i$ constitutes an orthonormal base, ρ's so-called eigen base. Density operators are used in quantum theory to describe the state of some system. If the system's state ρ is certain we call it a pure state and write $\rho = |s\rangle\langle s|$ for some unit ket $|s\rangle$. Systems whose state is uncertain are described by a mixed state $\rho = \sum_i p_i|s_i\rangle\langle s_i|$ which represents an ensemble of substates or pure states $\{(p_i, s_i)\}_i$ where the system is in substate s_i with probability p_i. Hence, the term "density" as in probability density.

It is possible to normalize a density matrix without committing to any particular decomposition. Only the trace function is required, because $norm(\rho) = \rho/Tr(\rho)$. Definition 2 mentions what the trace function does. However, notice that the same result is produced for any orthonormal base $\mathcal{B}$, including ρ's eigen base $\mathcal{B}_{eig} = \{|e_i\rangle\}_i$. Even though we do not know the content of $\mathcal{B}_{eig}$, we know that it exists. So we use it to show that dividing ρ by:

$$
\begin{aligned}
Tr(\rho) &= Tr(\sum_i p_i|e_i\rangle\langle e_i|) \\
&= \sum_j \langle e_j|(\sum_i p_i|e_i\rangle\langle e_i|)|e_j\rangle \\
&= \sum_{i,j} p_i\langle e_j|e_i\rangle\langle e_i|e_j\rangle \\
&= \sum_i p_i\langle e_i|e_i\rangle\langle e_i|e_i\rangle = \sum_i p_i
\end{aligned}
\tag{12}
$$

normalizes its probability distribution over eigen kets:

$$
\begin{aligned}
&\frac{\rho}{Tr(\rho)} = \frac{\sum_i p_i|e_i\rangle\langle e_i|}{\sum_j p_j} = \\
&\sum_i \frac{p_i}{\sum_j p_j}|e_i\rangle\langle e_i|
\end{aligned}
\tag{13}
$$

3 Semantic Space Model

We represent the meaning of words by density matrices. Specifically, a lexical item w is modeled as an ensemble $U_w = \{(p_i, u_i)\}_i$ of usages u_i and the corresponding probabilities p_i that w gets used "in the i-th manor". A word's usage is comprised of distributional information about its syntactic and semantic preferences, in the form of a ket $|u_i\rangle$. The density matrix $\rho_w = \sum_i p_i|u_i\rangle\langle u_i|$ represents the ensemble U_w. This section explains our method of extracting lexical density matrices from a dependency-parsed corpus. Once density matrices have been learned, we can predict the expected usage similarity of two words as a simple function of their density matrices. Our explication will be formally precise, but at the same time illustrate each principle through a toy example.

3.1 Dependency Hilbert Space

Our model learns the meaning of words from a dependency-parsed corpus. Our experiments have used the Stanford parser (de Marneffe and Manning, 2008), however any other dependency parser with broadly similar output could be used instead. A word's usage is learned from the type of dependency relations it has with its immediate neighbors in dependency graphs. Its semantic content is thus approximated by its "neighborhood", i.e., its co-occurrence frequency with neighboring words.

Neighborhoods are defined by a vocabulary $V = \{w_1, ..., w_{n_V}\}$ of the n_V most frequent (non-stop) words in the corpus. Let $Rel = \{\texttt{sub}^{-1}, \texttt{dobj}^{-1}, \texttt{amod}, \texttt{num}, \texttt{poss}, ...\}$ denote

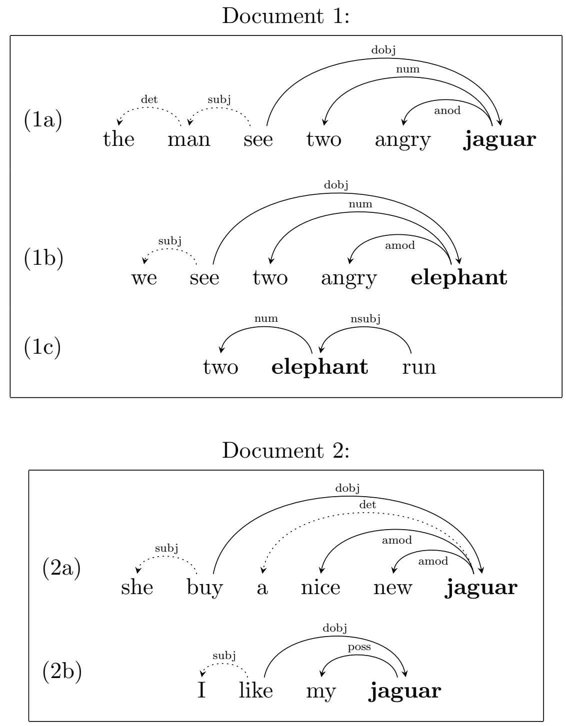

Figure 1: Example dependency trees in a toy corpus. Dotted arcs are ignored because they are either not connected to the target words *jaguar* and *elephant* or because their relation is not taken into account in constructing the semantic space. Words are shown as lemmas.

a subset of all dependency relations provided by the parser and their inverses. The choice of Rel is a model parameter. We considered only the most frequently occuring relations above a certain threshold, which turned out to be about half of the full inventory. Relation symbols with the superscript "$^{-1}$" indicate the inversion of the dependency direction (dependent to head). All other relation symbols have the conventional direction (head to dependent). Hence, $w \overset{\texttt{xyz}}{\rightarrow} v$ is equivalent to $v \overset{\texttt{xyz}^{-1}}{\rightarrow} w$. We then partition Rel into disjoint clusters of syntactically similar relations $Part = \{RC_1, ..., RC_{n_{Part}}\}$. For example, we consider syntactically similar relations which connect target words with neighbors with the same part of speech. Each relation cluster RC_k is assigned a Hilbert space $\mathcal{H}_k$ whose base kets $\{|w_j^{(k)}\rangle\}_j$ correspond to the words in $V = \{w_j\}_j$.

Figure 1 shows the dependency parses for a toy corpus consisting of two documents and five sentences. To create a density matrix for the target words *jaguar* and *elephant*, let us assume that we will consider the following relation clusters: $RC_1 = \{\texttt{dobj}^{-1}, \texttt{iobj}^{-1}, \texttt{agent}^{-1}, \texttt{nsubj}^{-1}, ...\}$, $RC_2 = \{\texttt{advmod}, \texttt{amod}, \texttt{tmod}, ...\}$ and $RC_3 = \{\texttt{nn}, \texttt{appos}, \texttt{num}, \texttt{poss}, ...\}$.

3.2 Mapping from Dependency Graphs to Kets

Next, we create kets which encode syntactic and semantic relations as follows. For each occurrence of the target word w in a dependency graph, we only consider the subtree made up of w and the immediate neighbors connected to it via a relation in Rel. In Figure 1, arcs from the dependency parse that we ignore are shown as dotted. Let the subtree of interest be $st = \{(RC_1, v_1), ..., (RC_{n_{Part}}, v_{n_{Part}})\}$, that is, w is connected to v_k via some relation in RC_k, for $k \in \{1, ..., n_{Part}\}$. For any relation cluster RC_k that does not feature in the subtree, let RC_k be paired with the abstract symbol $w_\emptyset$ in st. This symbol represents uncertainty about a potential RC_k-neighbor.

We convert all subtrees st in the corpus for the target word w into kets $|\psi_{st}\rangle \in \mathcal{H}_1 \otimes ... \otimes \mathcal{H}_{n_{Part}}$. These in turn make up the word's density matrix ρ_w. Before we do so, we assign each relation cluster RC_k a complex value $\alpha_k = e^{i\theta_k}$. The idea behind these values is to control for how much each subtree contributes to the overall density matrix. This becomes more apparent after we formulate our method of inducing usage kets and density matrices.

$$|\psi_{st}\rangle = \alpha_{st} \bigotimes_{(RC_k, v) \in st} |v\rangle_k, \tag{14}$$

where $\alpha_{st} = \sum_{(RC_k, v) \in st, v \neq w_\emptyset} \alpha_k$. Every RC_k paired with some neighbor $v \in V$ induces a basic subket $|v\rangle_k \in base(\mathcal{H}_k)$, i.e., a base ket of the k-th subspace or subsystem. All other subkets $|w_\emptyset\rangle_k = \sum_{v \in V} |V|^{-\frac{1}{2}} |v\rangle_k$ are in a uniformly weighted superposition of all base kets. The factor $|V|^{-\frac{1}{2}}$ ensures that $\langle w_\emptyset | w_\emptyset \rangle = 1$. The composed ket for the subtree st is again weighted by the complex-valued α_{st}.

α_{st} is the sum of complex values $\alpha_k = e^{i\theta_k}$, each with absolute value 1. Therefore, its own absolute value depends highly on the relative orientation θ_k among its summands: equal phases reinforce absolute value, but the more phases are opposed (i.e., their difference approaches π), the more they cancel out the sum's absolute value. Only those α_k contribute to this sum whose relation cluster is not paired with $w_\emptyset$. The choice of the parameters θ_k allows us to put more weight on some combinations of dependency relations than others.

(a)

	$\langle\psi_{st_1}\vert$	$\langle\psi_{st_2}\vert$	$\langle\psi_{st_3}\vert$
$\vert\psi_{st_1}\rangle$	$\neq 0$	$\neq 0$	$\neq 0$
$\vert\psi_{st_2}\rangle$	$\neq 0$	$\neq 0$	$\neq 0$
$\vert\psi_{st_3}\rangle$	$\neq 0$	$\neq 0$	$\neq 0$

(b)

	$\langle\psi_{st_1}\vert$	$\langle\psi_{st_2}\vert$	$\langle\psi_{st_3}\vert$
$\vert\psi_{st_1}\rangle$	$\neq 0$	0	0
$\vert\psi_{st_2}\rangle$	0	$\neq 0$	0
$\vert\psi_{st_3}\rangle$	0	0	$\neq 0$

Figure 2: Excerpts of density matrices that result from the dependency subtrees st_1, st_2, st_3. Element $m_{i,j}$ in row i and column j is $m_{i,j}|\psi_{st_i}\rangle\langle\psi_{st_j}|$ in Dirac notation. (a) All three subtrees are in the same document. Thus their kets contribute to diagonal and off-diagonal matrix elements. (b) Each subtree is in a separate document. Therefore their kets do not group, affecting only diagonal matrix elements.

Arbitrarily choosing $\theta_1 = \frac{\pi}{4}$, $\theta_2 = \frac{7\pi}{4}$, and $\theta_3 = \frac{3\pi}{4}$ renders the subtrees in Figure 1 as $|\psi_{st_{1a}}\rangle = e^{i\pi/4}|see\rangle_1|angry\rangle_2|two\rangle_3$, $|\psi_{st_{2a}}\rangle = \sqrt{2}|buy\rangle_1(|nice\rangle_2 + |new\rangle_2)|w_\emptyset\rangle_3$, $|\psi_{st_{2b}}\rangle = \sqrt{2}e^{i\pi/2}|like\rangle_1|w_\emptyset\rangle_2|my\rangle_3$, which are relevant for *jaguar*, and $|\psi_{st_{1b}}\rangle = e^{i\pi/4}|see\rangle_1|angry\rangle_2|two\rangle_3$, $|\psi_{st_{1c}}\rangle = \sqrt{2}e^{i\pi/2}|run\rangle_1|w_\emptyset\rangle_2|two\rangle_3$, which are relevant for *elephant*. The subscripts outside of the subkets correspond to those of the relation clusters RC_1, RC_2, RC_3 chosen in Section 3.1.

In sentence 2a, *jaguar* has two neighbors under RC_2. Therefore the subket from $\mathcal{H}_2$ is a superposition of the base kets $|nice\rangle_2$ and $|new\rangle_2$. This is a more intuitive formulation of the equivalent approach which first splits the subtree for *buy nice new jaguar* into two similar subtrees for *buy nice jaguar* and for *buy new jaguar*, and then processes them as seperate subtrees within the same document.

3.3 Creating Lexical Density Matrices

We assume that a word's usage is uniform throughout the same document. In our toy corpus in Figure 1, *jaguar* is always the direct object of the main verb. However, in Document 1 it is used in the animal sense, whereas in Document 2 it is used in the car sense. Even though the usage of *jaguar* in sentence (2b) is ambiguous, we group it with that of sentence (2a).

These considerations can all be comfortably en-

$$
\begin{aligned}
\rho_{jaguar} = {} & (|\psi_{st_{1a}}\rangle\langle\psi_{st_{1a}}| + (|\psi_{st_{2a}}\rangle + |\psi_{st_{2b}}\rangle)(\langle\psi_{st_{2a}}| + \langle\psi_{st_{2b}}|))/7 = \\
& 0.14|see\rangle_1|angry\rangle_2|two\rangle_3\langle see|_1\langle angry|_2\langle two|_3 + 0.29|buy\rangle_1|nice\rangle_2|w_\emptyset\rangle_3\langle buy|_1\langle nice|_2\langle w_\emptyset|_3 + \\
& 0.29|buy\rangle_1|nice\rangle_2|w_\emptyset\rangle_3\langle buy|_1\langle new|_2\langle w_\emptyset|_3 + 0.29|buy\rangle_1|new\rangle_2|w_\emptyset\rangle_3\langle buy|_1\langle nice|_2\langle w_\emptyset|_3 + \\
& 0.29|buy\rangle_1|new\rangle_2|w_\emptyset\rangle_3\langle buy|_1\langle new|_2\langle w_\emptyset|_3 + 0.29e^{\pi/2}|like\rangle_1|w_\emptyset\rangle_2|my\rangle_3\langle buy|_1\langle nice|_2\langle w_\emptyset|_3 + \\
& 0.29e^{\pi/2}|like\rangle_1|w_\emptyset\rangle_2|my\rangle_3\langle buy|_1\langle new|_2\langle w_\emptyset|_3 + 0.29e^{-\pi/2}|buy\rangle_1|nice\rangle_2|w_\emptyset\rangle_3\langle like|_1\langle w_\emptyset|_2\langle my|_3 + \\
& 0.29e^{-\pi/2}|buy\rangle_1|new\rangle_2|w_\emptyset\rangle_3\langle like|_1\langle w_\emptyset|_2\langle my|_3 + 0.29|like\rangle_1|w_\emptyset\rangle_2|my\rangle_3\langle like|_1\langle w_\emptyset|_2\langle my|_3
\end{aligned}
$$

$$
\begin{aligned}
\rho_{elephant} = {} & ((|\psi_{st_{1b}}\rangle + |\psi_{st_{1c}}\rangle)(\langle\psi_{st_{1b}}| + \langle\psi_{st_{1c}}|))/3 = \\
& 0.33|see\rangle_1|angry\rangle_2|two\rangle_3\langle see|_1\langle angry|_2\langle two|_3 + 0.47e^{\pi/4}|run\rangle_1|w_\emptyset\rangle_2|two\rangle_3\langle see|_1\langle angry|_2\langle two|_3 + \\
& 0.47e^{-\pi/4}|see\rangle_1|angry\rangle_2|two\rangle_3\langle run|_1\langle w_\emptyset|_2\langle two|_3 + 0.67|run\rangle_1|w_\emptyset\rangle_2|two\rangle_3\langle run|_1\langle w_\emptyset|_2\langle two|_3
\end{aligned}
$$

$$
\begin{aligned}
& Tr(\rho_{jaguar}\rho_{elephant}) = Tr(0.05|\psi_{st_{1a}}\rangle\langle\psi_{st_{1b}}| + 0.05e^{i\pi/4}|\psi_{st_{1a}}\rangle\langle\psi_{st_{1c}}|) = Tr(0.05|see\rangle_1|angry\rangle_2|two\rangle_3\langle see|_1\langle angry|_2\langle two|_3 + \\
& 0.07e^{-\pi/4}|see\rangle_1|angry\rangle_2|two\rangle_3\langle run|_1\langle w_\emptyset|_2\langle two|_3) = \sum_{|b\rangle \in base(\mathcal{H}_1\otimes\mathcal{H}_2\otimes\mathcal{H}_3)} \langle b|(0.05|see\rangle_1|angry\rangle_2|two\rangle_3\langle see|_1\langle angry|_2\langle two|_3 + \\
& 0.07e^{-\pi/4}|see\rangle_1|angry\rangle_2|two\rangle_3\langle run|_1\langle w_\emptyset|_2\langle two|_3)|b\rangle = 0.05
\end{aligned}
$$

Figure 3: Lexical density matrices for the words *jaguar* and *elephant* and their similarity.

coded in a density matrix. This is simply generated via the outer product of our subtree kets $|\psi_{st}\rangle$. For example, $\rho_{D_1,jaguar} = |\psi_{st_{1a}}\rangle\langle\psi_{st_{1a}}|$ represents the contribution that document D_1 makes to ρ_{jaguar}. Document D_2, however, has more than one ket relevant to ρ_{jaguar}. Due to our assumption of document-internal uniformity of word usage, we group D_2's subtree-kets additively: $\rho_{D_2,jaguar} = (|\psi_{st_{2a}}\rangle + |\psi_{st_{2b}}\rangle)(\langle\psi_{st_{2a}}| + \langle\psi_{st_{2b}}|)$. The target word's density matrix ρ_w is the normalized sum of all density matrices $\rho_{D,w}$ obtained from each D:

$$
\rho_{D,w} = \left(\sum_{st \in ST_{D,w}} |\psi_{st}\rangle\right)\left(\sum_{st \in ST_{D,w}} \langle\psi_{st}|\right) \tag{15}
$$

where $ST_{D,w}$ is the set of all subtrees for target word w in document D. To illustrate the difference that this grouping makes, consider the density matrices in Figure 2. Whereas in (a) the subtrees st_1, st_2, st_3 share a document, in (b) they are from distinct documents. This grouping causes them to not only contribute to diagonal matrix elements, e.g., $|\psi_{st_2}\rangle\langle\psi_{st_2}|$, as in (b), but also to off diagonal ones, e.g., $|\psi_{st_2}\rangle\langle\psi_{st_1}|$, as in (a).

Over the course of many documents the summation of all contributions, no matter how small or large the groups are, causes "clusters of weight" to form, which hopefully coincide with word usages. As mentioned in Section 3.2, adding complex-valued matrix elements increases or decreases the sum's absolute value depending on relative phase orientation. This makes it possible for interference to occur. Since the same word appears in varying contexts, the corresponding complex-valued outer products interact upon summation. Finally, the density matrix gets normalized, i.e., divided by its trace. This leaves the distributional information intact and merely normalizes the probabilities. Figure 3 illustrates the estimation of the density matrices for the words *jaguar* and *elephant* from the toy corpus in Figure 1.

3.4 Usage Similarity

Decomposing the density matrix of the target word w, $\rho_w = \sum_i p_i |u_i\rangle\langle u_i|$ recovers the usage ensemble $U_w = \{(p_i, u_i)\}_i$. However, in general there are infinitely many possible ensembles which ρ_w might represent. This subsection explains our metric for estimating the usage similarity of two words. The math involved shows that we can avoid the question of how to best decompose ρ_w.

We compute the usage similarity of two words w and v by comparing each usage of w with each usage of v and weighting these similarity values with the corresponding usage probabilities. Let $\rho_w = \sum_i p_i^{(w)} |u_i^{(w)}\rangle\langle u_i^{(w)}|$ and $\rho_v = \sum_i p_i^{(v)} |u_i^{(v)}\rangle\langle u_i^{(v)}|$. The similarity of some usage kets $|u_i^{(w)}\rangle$ and $|u_j^{(v)}\rangle$ is obtained, as is common in the literature, by their inner product $\langle u_i^{(w)}|u_j^{(v)}\rangle$. However, as this is a complex value, we multiply it with its complex conjugate, rendering the real value $\langle u_j^{(v)}|u_i^{(w)}\rangle\langle u_i^{(w)}|u_j^{(v)}\rangle = |\langle u_i^{(w)}|u_j^{(v)}\rangle|^2$. Therefore, in total the expected similarity of w and v is:

$$sim(w,v) = \sum_{i,j} p_i^{(w)} p_j^{(v)} \langle u_j^{(v)} | u_i^{(w)} \rangle \langle u_i^{(w)} | u_j^{(v)} \rangle \quad (16)$$

$$= Tr\left(\sum_{i,j} p_i^{(w)} p_j^{(v)} | u_i^{(w)} \rangle \langle u_i^{(w)} | u_j^{(v)} \rangle \langle u_j^{(v)} | \right) =$$

$$Tr\left((\sum_i p_i^{(w)} | u_i^{(w)} \rangle \langle u_i^{(w)} |)(\sum_j p_j^{(v)} | u_j^{(v)} \rangle \langle u_j^{(v)} |) \right)$$

$$= Tr(\rho_w \rho_v)$$

We see that the similarity function simply reduces to multiplying ρ_w with ρ_v and applying the trace function. The so-called cyclic property of the trace function (i.e., $Tr(M_1M_2) = Tr(M_2M_1)$ for any two matrices M_1, M_2) gives us the corollary that this particular similarity function is symmetric.

Figure 3 (bottom) shows how to calculate the similarity of *jaguar* and *elephant*. Only the coefficient of the first outer product survives the tracing process because its ket and bra are equal modulo transpose conjugate. As for the second outer product, $0.05e^{i\pi/4} \langle b | \psi_{st_{1a}} \rangle \langle \psi_{st_{1c}} | b \rangle$ is 0 for all base kets $|b\rangle$.

3.5 What Does This Achieve?

We represent word meanings as described above for several reasons. The density matrix decomposes into usages each of which are a superposition of combinations of dependents. Internally, these usages are established automatically by way of "clustering".

Our model is parameterized with regard to the phases of sub-systems (i.e., clusters of syntactic relations) which allows us to make optimal use of interference, as this plays a large role in the overall quality of representation. It is possible for a combination of (groups of) dependents to get entangled if they repeatedly appear together under the same word, and only in that combination. If the co-occurence of (groups of) dependents is uncorrelated, though, they remain unentangled. Quantum entanglement gives our semantic structures the potential for long-distance effects, once quantum measurement becomes involved. This is in analogy to the nonlocal correlation between properties of subatomic particles, such as the magnetic spin of electrons or the polarization of photons. Such an extension to our implementation will also uncover which sets of measurements are order-sensitive, i.e., incompatible.

Our similarity metric allows two words to "select" each other's usages via their pairwise inner products. Usage pairs with a high distributional similarity roughly "align" and then get weighted by the probabilities of those usages. Two words are similar if they are substitutable, that is, if they can be used in the same syntactic environment and have the same meaning. Hopefully, this leads to more accurate estimation of distributional similarity and can be used to compute word meaning in context.

4 Experimental Setup

Data All our experiments used a dependency parsed and lemmatized version of the British National Corpus (BNC). As mentioned in Section 3, we obtained dependencies from the output of the Stanford parser (de Marneffe and Manning, 2008). The BNC comprises 4,049 texts totalling approximately 100 million words.

Evaluation Tasks We evaluated our model on word similarity and association. Both tasks are employed routinely to assess how well semantic models predict human judgments of word relatedness. We used the WordSim353 test collection (Finkelstein et al., 2002) which consists of similarity judgments for word pairs. Participants gave each pair a similarity rating using a 0 to 10 scale (e.g., *tiger–cat* are very similar, whereas *delay–racism* are not). The average rating for each pair represents an estimate of the perceived similarity of the two words. The collection contains ratings for 437 unique words (353 pairs) all of which appeared in our corpus. Word association is a slightly different task: Participants are given a cue word (e.g., *rice*) and asked to name an associate in response (e.g., *Chinese, wedding, food, white*). We used the norms collected by Nelson et al. (1998). We estimated the strength of association between a cue and its associate, as the relative frequency with which it was named. The norms contain 9,968 unique words (70,739 pairs) out of which 9,862 were found in our corpus, excluding multiword expressions.

For both tasks, we used correlation analysis to examine the degree of linear relationship between human ratings and model similarity values. We report correlation coefficients using Spearman's rank correlation coefficient.

Quantum Model Parameters The quantum framework presented in Section 3 is quite flexible. Depending on the choice of dependency relations *Rel*, dependency clusters RC_j, and complex values $\alpha_j = e^{i\theta_j}$, different classes of models can be derived. To explore these parameters, we partitioned the WordSim353 dataset and Nelson et al.'s (1998) norms into a development and test set following a 70–30 split. We tested 9 different intuitively chosen relation partitions $\{RC_1, ..., RC_{n_{Part}}\}$, creating models that considered only neighboring heads, models that considered only neighboring dependents, and models that considered both. For the latter two we experimented with partitions of one, two or three clusters. In addition to these more coarse grained clusters, for models that included both heads and dependents we explored a partition with twelve clusters broadly corresponding to objects, subjects, modifiers, auxiliaries, determiners and so on. In all cases stopwords were not taken into account in the construction of the semantic space.

For each model variant we performed a grid search over the possible phases $\theta_j = k\pi$ with range $k = \frac{0}{4}, \frac{1}{4}, ..., \frac{7}{4}$ for the complex-valued α_j assigned to the respective relation cluster RC_j (see Section 3.2 for details). In general, we observed that the choice of dependency relations and their clustering as well as the phases assigned to each cluster greatly influenced the semantic space. On both tasks, the best performing model had the relation partition described in Section 3.1. Section 5 reports our results on the test set using this model.

Comparison Models We compared our quantum space against three classical distributional models. These include a simple semantic space, where a word's meaning is a vector of co-occurrences with neighboring words (Mitchell and Lapata, 2010), a syntax-aware space based on weighted distributional triples that encode typed co-occurrence relations among words (Baroni and Lenci, 2010) and word embeddings computed with a neural language model (Bengio, 2001; Collobert and Weston, 2008) For all three models we used parameters that have been reported in the literature as optimal.

Specifically, for the simple co-occurrence-based space we follow the settings of Mitchell and Lapata (2010): a context window of five words on either side of the target word and 2,000 vector dimensions (i.e., the 2000 most common context words in the BNC). Vector components were set to the ratio of the probability of the context word given the target word to the probability of the context word overall. For the neural language model, we adopted the best performing parameters from our earlier comparison of different vector sources for distributional semantics (Blacoe and Lapata, 2012) where we also used the BNC for training. There we obtained best results with 50 dimensions, a context window of size 4, and an embedding learning rate of 10^{-9}. Our third comparison model uses Baroni and Lenci's (2010) third-order tensor[2] which they obtained from a very large dependency-parsed corpus containing approximately 2.3 billion words. Their tensor assigns a mutual information score to instances of word pairs w, v and a linking word l. We obtained vectors $\overrightarrow{w}$ from the tensor following the methodology proposed in Blacoe and Lapata (2012) using 100 (l, v) contexts as dimensions.

Models	WordSim353	Nelson Norms
SDS	0.433	0.151
DM	0.318	0.123
NLM	0.196	0.091
QM	**0.535**	**0.185**

Table 1: Performance of distributional models on WordSim353 dataset and Nelson et al.'s (1998) norms (test set). Correlation coefficients are all statistically significant ($p < 0.01$).

5 Results

Our results are summarized in Table 1. As can be seen, the quantum model (QM) obtains performance superior to other better-known models such as Mitchell and Lapata's (2010) simple semantic space (SDS), Baroni and Lenci's (2010) distributional memory tensor (DM), and Collobert and Weston's (2008) neural language model (NLM). Our results on the association norms are comparable to the state of the art (Silberer and Lapata, 2012; Griffiths et al., 2007). With regard to WordSim353, Huang et al. (2012) report correlations in the range of 0.713–0.769, however they use Wikipedia as a training corpus and a more sophisticated version of the NLM presented here, that takes into account global context and performs word sense discrimination. In the future, we also plan to evaluate our model on larger Wikipedia-scale corpora. We would also like to model semantic composition as our approach can do this easily by taking advantage of the notion of quantum measurement. Specifically, we

[2] Available at `http://clic.cimec.unitn.it/dm/`.

Models	bar	order
SDS	pub, snack, restaurant, grill, coctail	form, direct, procedure, plan, request
DM	counter, rack, strip, pipe, code	court, demand, form, law, list
NLM	room, pole, drink, rail, coctail	direct, command, plan, court, demand
QM	prison, liquor, beer, club, graph	organization, food, law, structure, regulation
HS	drink, beer, stool, alcohol, grill	food, form, law, heat, court

Table 2: Associates for *bar* and *order* ranked according to similarity. Underlined associates overlap with the human responses (HS).

can work out the meaning of a dependency tree by measuring the meaning of its heads in the context of their dependents.

Table 2 shows the five most similar associates (ordered from high to low) for the cues *bar* and *order* for the quantum model and the comparison models. We also show the human responses (HS) according to Nelson et al.'s (1998) norms. The associates generated by the quantum model correspond to several different meanings correlated with the target. For example, *prison* refers to the "behind bars" sense of *bar*, *liquor* and *beer* refer to what is consumed or served in bars, *club* refers to the entertainment function of bars, whereas *graph* refers to how data is displayed in a chart.

6 Related Work

Within cognitive science the formal apparatus of quantum theory has been used to formulate models of cognition that are superior to those based on traditional probability theory. For example, conjunction fallacies[3] (Tversky and Kahneman, 1983) have been explained by making reference to quantum theory's context dependence of the probability assessment. Violations of the sure-thing principle[4] (Tversky and Shafir, 1992) have been modeled in terms of a quantum interference effect. And the asymmetry of similarity relations has been explained by postulating that different concepts correspond to subspaces of different dimensionality (Pothos and Busemeyer, 2012). Several approaches have drawn on quantum theory in order to model semantic phenomena such as concept combination (Bruza and Cole, 2005), the emergence of new concepts (Aerts and Gabora, 2005), and the human mental lexicon (Bruza et al., 2009). Chen (2002) captures syllogisms in a quantum theoretic framework; the model takes statements like *All whales are mammals and all mammals are animals* as input and outputs conclusions like *All whales are animals*.

The first attempts to connect the mathematical basis of semantic space models with quantum theory are due to Aerts and Czachor (2004) and Bruza and Cole (2005). They respectively demonstrate that Latent Semantic Analysis (Landauer and Dumais, 1997) and the Hyperspace Analog to Language model (Lund and Burgess, 1996) are essentially Hilbert space formalisms, without, however, providing concrete ways of building these models beyond a few hand-picked examples. Interestingly, Bruza and Cole (2005) show how lexical operators may be contrived from corpus co-occurrence counts, albeit admitting to the fact that their operators do not provide sensical eigenkets, most likely because of the simplified method of populating the matrix from corpus statistics. Grefenstette et al. (2011) present a model for capturing semantic composition in a quantum theoretical context, although it appears to be reducible to the classical probabilistic paradigm. It does not make use of the unique aspects of quantum theory (e.g., entanglement, interference, or quantum collapse).

Our own work follows Aerts and Czachor (2004) and Bruza and Cole (2005) in formulating a model that exhibits important aspects of quantum theory. Contrary to them, we present a fully-fledged semantic space rather than a proof-of-concept. We obtain quantum states (i.e., lexical representations) for each word by taking its *syntactic* context into account. Quantum states are expressed as density operators rather than kets. While a ket can only capture one pure state of a system, a density operator contains an ensemble of pure states which we argue is advantageous from a modeling perspective. Within this framework, not only can we compute the meaning of individual words but also phrases or sentences, without postulating any additional operations. Compositional meaning reduces to quantum measurement at each inner node of the (dependency) parse of the structure in question.

[3] A conjunction fallacy occurs when it is assumed that specific conditions are more probable than a single general one.

[4] The principle is the expectation that human behavior ought to conform to the law of total probability

References

Diederik Aerts and Marek Czachor. 2004. Quantum aspects of semantic analysis and symbolic artificial intelligence. *Journal of PhysicsA: Mathematical and General*, 37:123–132.

Diederik Aerts and Liane Gabora. 2005. A state-context-property model of concepts and their combinations ii: A hilbert space representation. *Kybernetes*, 1–2(34):192–221.

Diederik Aerts. 2009. Quantum structure in cognition. *Journal of Mathematical Psychology*, 53:314–348.

Belal E. Baaquie. 2004. *Quantum Finance: Path Integrals and Hamiltonians for Oprions and Interest Rates*. Cambridge University Press, Cambridge.

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

Yoshua Bengio. 2001. Neural net language models. *Scholarpedia*, 3(1):3881.

William Blacoe and Mirella Lapata. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 546–556, Jeju Island, Korea, July. Association for Computational Linguistics.

Peter D. Bruza and Richard J. Cole. 2005. Quantum logic of semantic space: An exploratory investigation of context effects in practical reasoning. In S. Artemov, H. Barringer, S. A. d'Avila Garcez, L. C. Lamb, and J. Woods, editors, *We Will Show Them: Essays in Honour of Dov Gabbay*, volume 1, pages 339–361. London: College Publications.

Peter D. Bruza, Kirsty Kitto, Douglas McEnvoy, and Cathy McEnvoy. 2008. Entangling words and meaning. In *Second Quantum Interaction Symposium*. Oxford University.

Peter D. Bruza, Kirsty Kitto, Douglas Nelson, and Cathy McEvoy. 2009. Is there something quantum-like in the human mental lexicon? *Journal of Mathematical Psychology*, 53(5):362–377.

Jerome R. Busemeyer and Peter D. Bruza. 2012. *Quantum Models of Cognition and Decision*. Cambridge University Press, Cambridge.

Joseph C. H. Chen. 2002. *Quantum computation and natural language processing*. Ph.D. thesis, Universität Hamburg.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multita sk learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167, New York, NY. ACM.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK.

Guy Denhière and Benoît Lemaire. 2004. A computational model of children's semantic memory. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, pages 297–302, Mahwah, NJ. Lawrence Erlbaum Associates.

Paul A. M. Dirac. 1939. A new notation for quantum mechanics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 35:416–418.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1):116–131, January.

John R. Firth. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pages 1–32. Philological Society, Oxford.

Edward Grefenstette, Mehrnoosh Sadrzadeh, Stephen Clark, Bob Coecke, and Stephen Pulman. 2011. Concrete sentence spaces for compositional distributional models of meaning. *Proceedings of the 9th International Conference on Computational Semantics (IWCS11)*, pages 125–134.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114(2):211–244.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea, July. Association for Computational Linguistics.

R. I. G. Hughes. 1989. *The Structure and Interpretation of Quantum Mechnics*. Harvard University Press, Cambridge, MA.

Chris J. Isham. 1989. *Lectures on Quantum Theory*. Singapore: World Scientific.

Andrei Y. Khrennikov. 2010. *Ubiquitous Quantum Structure: From Psychology to Finance*. Springer.

Daniel Kleppner and Roman Jackiw. 2000. One hundred years of quantum physics. *Science*, 289(5481):893–898.

Darrell R. Laham. 2000. *Automated Content Assessment of Text Using Latent Semantic Analysis to Simulate Human Cognition*. Ph.D. thesis, University of Colorado at Boulder.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the joint Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, pages 768–774, Montréal, Canada.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, 28:203–208.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 38(8):1388–1429.

Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. 1998. The University of South Florida Word Association, Rhyme, and Word Fragment Norms.

Michael A. Nielsen and Isaac L. Chuang. 2010. *Quantum Computation and Information Theory*. Cambridge University Press, Cambridge.

Daniel Osherson and Edward E. Smith. 1981. On the adequacy of prototype theory as a theory of concepts. *Cognition*, 9:35–38.

Emmanuel M. Pothos and Jerome R. Busemeyer. 2012. Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences*. to appear.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, Jeju Island, Korea. Association for Computational Linguistics.

Richard Socher, Eric H. Huang, Jeffrey Pennin, Andrew Y. Ng, and Christopher D. Manning. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 801–809.

Amos Tversky and Daniel Kahneman. 1983. Extensional versus intuitive reasoning: The conjuctive fallacy in probability judgment. *Psychological Review*, 4(90):293–315.

Amos Tversky and Eldar Shafir. 1992. The disjunction effect in choice under uncertainty. *Psychological Science*, 3(5):305–309.

Vlatko Vedral. 2006. *Introduction to Quantum Information Science*. Oxford University Press, New York.