Graded hyponymy for compositional distributional semantics

Dea Bankova[1], *Bob Coecke*[1], *Martha Lewis*[2], *and Dan Marsden*[1]
[1] Quantum Group, University of Oxford
[2] ILLC, University of Amsterdam

ABSTRACT

The categorical compositional distributional model of natural language provides a conceptually motivated procedure to compute the meaning of a sentence, given its grammatical structure and the meanings of its words. This approach has outperformed other models in mainstream empirical language processing tasks, but lacks an effective model of lexical entailment. We address this shortcoming by exploiting the freedom in our abstract categorical framework to change our choice of semantic model. This allows us to describe hyponymy as a graded order on meanings, using models of partial information used in quantum computation. Quantum logic embeds in this graded order.

Keywords: categorical compositional distributional semantics, computational linguistics, entailment, density operator

1 INTRODUCTION

Finding a formalization of language in which the meaning of a sentence can be computed from the meaning of its parts has been a long-standing goal in formal and computational linguistics.

Distributional semantics represents individual word meanings as vectors in finite dimensional real vector spaces. On the other hand, symbolic accounts of meaning combine words via compositional rules to form phrases and sentences. These two approaches are in some sense orthogonal. Distributional schemes have no obvious compositional structure, whereas compositional models lack a canonical way of determining the meaning of individual words. In Coecke *et al.* (2010), the authors develop the categorical compositional distributional model of natural language semantics. This model exploits the

shared categorical structure of pregroup grammars and vector spaces to provide a compositional structure for distributional semantics. It has produced state-of-the-art results in measuring sentence similarity (Kartsaklis *et al.* 2012; Grefenstette and Sadrzadeh 2011), effectively describing aspects of the human understanding of sentences.

A satisfactory account of natural language should incorporate a suitable notion of lexical entailment. Until recently, categorical compositional distributional models of meaning have lacked this crucial feature. In order to address the entailment problem, we exploit the freedom inherent in our abstract categorical framework to change models. We move from a pure state setting to a category used to describe mixed states and partial knowledge in the semantics of categorical quantum mechanics. Meanings are now represented by density matrices rather than simple vectors. We use this extra flexibility to capture the concept of hyponymy, where one word may be seen as an instance of another. For example, *red* is a hyponym of *colour*. The hyponymy relation can be associated with a notion of logical entailment. Some entailment is crisp, for example: *dog* entails *animal*. However, we may also wish to permit entailments of differing strengths. For example, the concept *dog* gives high support to the concept *pet*, but does not completely entail it: some dogs are working dogs. The hyponymy relation we describe here can account for these phenomena. Some crisp entailment can be seen as encoding linguistic knowledge. The kind of entailment we are interested in here is, in general, about the properties that objects have in the world, rather than grammatically based entailment. In particular, we explicitly avoid downward-monotone contexts such as negation. We do, however, examine the hyponymy between an adjective-noun compound and the head noun. We should also be able to measure entailment strengths at the sentence level. For example, we require that *Cujo is a dog* crisply entails *Cujo is an animal*, but that the statement *Cujo is a dog* does not completely entail *Cujo is a pet*. Again, the relation we describe here will successfully describe this behaviour at the sentence level. Closely related to the current work are the ideas in Balkır (2014), Balkır *et al.* (2016), and Sadrzadeh *et al.* (2018). In this work, the authors develop a graded form of entailment based on von Neumann entropy and with links to the distributional inclusion hypotheses developed by Geffet and Dagan (2005). The authors

show how entailment at the word level carries through to entailment at the sentence level. However, this is done without taking account of the grading. In contrast, the measure that we develop here provides a lower bound for the entailment strength between sentences, based on the entailment strength between words. Some of the work presented here was developed in the first author's MSc thesis (Bankova 2015).

An obvious choice for a logic built upon vector spaces is quantum logic (Birkhoff and von Neumann 1936). Briefly, this logic represents propositions about quantum systems as projection operators on an appropriate Hilbert space. These projections form an orthomodular lattice where the distributive law fails in general. The logical structure is then inherited from the lattice structure in the usual way. In the current work, we propose an order that embeds the orthomodular lattice of projections, and so contains quantum logic. This order is based on the Löwner ordering with propositions represented by density matrices. When this ordering is applied to density matrices with the standard trace normalization, no propositions compare, and therefore the Löwner ordering is useless as applied to density operators. The trick we use is to develop an approximate entailment relationship which arises naturally from any commutative monoid. We introduce this in general terms and describe conditions under which this gives a graded measure of entailment. This grading becomes continuous with respect to noise. Our framework is flexible enough to subsume the Bayesian partial ordering of Coecke and Martin (2011) and provides it with a grading. A procedure is given for determining the hyponymy strength between *any* pair of phrases of the same overall grammatical type. The pair of phrases can have differing lengths. So, for example, we can compare 'blond men' to 'men', as these are both noun phrases. This is possible because within categorical compositional semantics, phrases of each type are reduced to one common space according to their type, and can be compared within that space. Furthermore, this notion is consistent with hyponymy at the word level, giving a lower bound on phrase hyponymy.

Density matrices have also been used in other areas of distributional semantics such as Kartsaklis (2015), Piedeleu (2014),

Piedeleu *et al.* (2015), and Blacoe *et al.* (2013). Quantum logic is used in (Widdows and Peters 2003) and Rijsbergen (2004).

Entailment is an important and thriving area of research within distributional semantics. The PASCAL Recognising Textual Entailment Challenge (Dagan *et al.* 2006) has attracted a large number of researchers in the area and generated a number of approaches. Previous lines of research on entailment for distributional semantics investigate the development of directed similarity measures which can characterize entailment (Weeds *et al.* 2004; Kotlerman *et al.* 2010; Lenci and Benotto 2012). Geffet and Dagan (2005) introduce a pair of *distributional inclusion hypotheses*, where if a word v entails another word w, then all the typical features of the word v will also occur with the word w. Conversely, if all the typical features of v also occur with w, v is expected to entail w. Clarke (2009) defines a vector lattice for word vectors, and a notion of graded entailment with the properties of a conditional probability. Rimell (2014) explores the limitations of the distributional inclusion hypothesis by examining the properties of those features that are not shared between words. An interesting approach in Kiela *et al.* (2015) is to incorporate other modes of input into the representation of a word. Measures of entailment are based on the dispersion of a word representation, together with a similarity measure. All of these look at entailment at the word level.

Attempts have also been made to incorporate entailment measures with elements of compositionality. Baroni *et al.* (2012) exploit the entailment relations between adjective-noun and noun pairs to train a classifier that can detect similar relations. They further develop a theory of entailment for quantifiers. The approach that we propose here has the characteristic that it can be applied to more types of phrases and sentences than just adjective-noun and noun-noun type phrases.

Another approach to compositional vector-based entailment is the use of deep neural networks to represent logical semantics, as in Bowman *et al.* (2015), for example. The drawback with the use of this sort of method is that the transparency of the compositional method is lost: the networks may indeed learn how to represent logical semantics but it is not clear how they do so. In contrast, the method we propose has a clear basis in formal semantics and links to quantum logic.

2 CATEGORICAL COMPOSITIONAL DISTRIBUTIONAL MEANING

Compositional and distributional accounts of meaning are unified in Coecke *et al.* (2010), constructing the meaning of sentences from the meanings of their component parts using their syntactic structure.

2.1 *Pregroup grammars*

In order to describe syntactic structure, we use Lambek's pregroup grammars (Lambek 1997). Within the standard categorical compositional distributional model, it is possible to move to other forms of categorial grammar, as argued in Coecke *et al.* (2013). This is due to the fact that the category of finite-dimensional vector spaces is particularly well-behaved, and so grammars with greater or lesser structure may be used. A pregroup $(P, \leq, \cdot, 1, (-)^l, (-)^r)$ is a partially ordered monoid $(P, \leq, \cdot, 1)$ where each element $p \in P$ has a left adjoint p^l and a right adjoint p^r, such that the following inequalities hold:

$$p^l \cdot p \leq 1 \leq p \cdot p^l \quad \text{and} \quad p \cdot p^r \leq 1 \leq p^r \cdot p \tag{1}$$

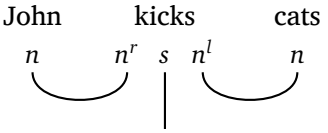
Intuitively, we think of the elements of a pregroup as linguistic types. The monoidal structure allows us to form composite types, and the partial order encodes type reduction. The important right and left adjoints then enable the introduction of types requiring further elements on either their left or right respectively.

The pregroup grammar $\mathsf{Preg}_{\mathscr{B}}$ over an alphabet $\mathscr{B}$ is freely constructed from the atomic types in $\mathscr{B}$. In what follows we use an alphabet $\mathscr{B} = \{n, s\}$. We use the type s to denote a declarative sentence and n to denote a noun. A transitive verb can then be denoted $n^r s n^l$. If a string of words and their types reduces to the type s, the sentence is judged grammatical. The sentence *John kicks cats* is typed $n\ (n^r s n^l)\ n$, and can be reduced to s as follows:

$$n\ (n^r s n^l)\ n \leq 1 \cdot s n^l n \leq 1 \cdot s \cdot 1 \leq s$$

This symbolic reduction can also be expressed graphically, as shown in Figure 1. In this diagrammatic notation, the elimination of types by means of the inequalities $n \cdot n^r \leq 1$ and $n^l \cdot n \leq 1$ is denoted by a 'cup'. The fact that the type s is retained is represented by a straight wire.

Figure 1:
A transitive sentence in the graphical calculus



2.2 *Compositional distributional models*

The symbolic account and distributional approaches are linked by the fact that they are both compact closed categories. This compatibility allows the compositional rules of the grammar to be applied in the vector space model. In this way, we can map syntactically well-formed strings of words into one shared meaning space.

A *compact closed category* is a monoidal category in which for each object A there are left and right dual objects A^l and A^r, and corresponding unit and counit morphisms $\eta^l : I \to A \otimes A^l$, $\eta^r : I \to A^r \otimes A$, $\epsilon^l : A^l \otimes A \to I$, $\epsilon^r : A \otimes A^r \to I$ such that the *snake equations* hold:

$$(1_A \otimes \epsilon^l) \circ (\eta^l \otimes 1_A) = 1_A \qquad (\epsilon^r \otimes 1_A) \circ (1_A \otimes \eta^r) = 1_A$$

$$(\epsilon^l \otimes 1_{A^l}) \circ (1_{A^l} \otimes \eta^l) = 1_{A^l} \qquad (1_{A^r} \otimes \epsilon^r) \circ (\eta^r \otimes 1_{A^r}) = 1_{A^r}$$

The underlying poset of a pregroup can be viewed as a compact closed category with the monoidal structure given by the pregroup monoid, and $\epsilon^l, \eta^l, \eta^r, \epsilon^r$ the unique morphisms witnessing the inequalities of (1).

Distributional vector space models live in the category **FHilb** of finite dimensional real Hilbert spaces and linear maps. **FHilb** is compact closed. Each object V is its own dual and the left and right unit and counit morphisms coincide. Given a fixed basis $\{|v_i\rangle\}_i$ of V, we define the unit by $\eta : \mathbb{R} \to V \otimes V :: 1 \mapsto \sum_i |v_i\rangle \otimes |v_i\rangle$ and counit by $\epsilon : V \otimes V \to \mathbb{R} :: \sum_{ij} c_{ij} |v_i\rangle \otimes |v_j\rangle \mapsto \sum_i c_{ii}$. Here, we use the physicists' bra-ket notation, for details see Nielsen and Chuang (2011).

2.3 *Graphical calculus*

The morphisms of compact closed categories can be expressed in a convenient graphical calculus (Kelly and Laplaza 1980) which we will exploit in the following sections. Objects are labelled wires, and morphisms are given as vertices with input and output wires. Composing morphisms consists of connecting input and output wires, and the tensor product is formed by juxtaposition, as shown in Figure 2.

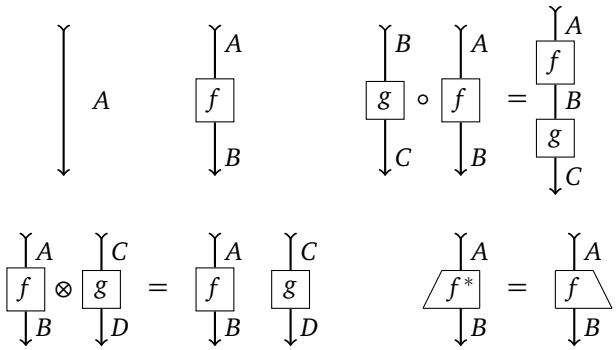


Figure 2:
Monoidal graphical calculus

By convention the wire for the monoidal unit is omitted. The morphisms ϵ and η can then be represented by 'cups' and 'caps' as shown in Figure 3. The snake equations can be seen as straightening wires, as shown in Figure 4.

Figure 3:
Compact structure graphically

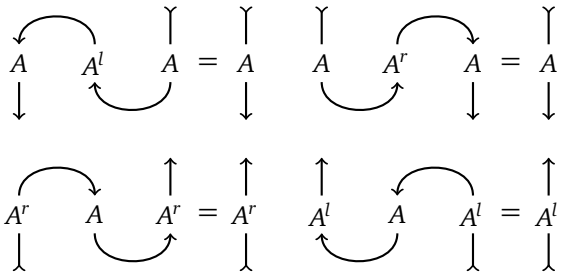


Figure 4:
The snake equations

2.4 *Grammatical Reductions in Vector Spaces*

Following Preller and Sadrzadeh (2011), reductions of the pregroup grammar may be mapped onto the category **FHilb** of finite dimensional Hilbert spaces and linear maps using an appropriate strong monoidal functor Q:

$$\mathsf{Q} : \mathbf{Preg} \rightarrow \mathbf{FHilb}$$

Strong monoidal functors automatically preserve the compact closed structure. For our example $\mathsf{Preg}_{\{n,s\}}$, we must map the noun and

sentence types to appropriate finite dimensional vector spaces:

$$\mathrm{Q}(n) = N \qquad \mathrm{Q}(s) = S$$

Composite types are then constructed functorially using the corresponding structure in **FHilb**. Each morphism α in the pregroup is mapped to a linear map interpreting sentences of that grammatical type. Then, given word vectors $|w_i\rangle$ with types p_i, and a type reduction $\alpha : p_1, p_2, \ldots, p_n \to s$, the meaning of the sentence $w_1 w_2 \ldots w_n$ is given by:

$$|w_1 w_2 \ldots w_n\rangle = \mathrm{Q}(\alpha)(|w_1\rangle \otimes |w_2\rangle \otimes \ldots \otimes |w_n\rangle)$$

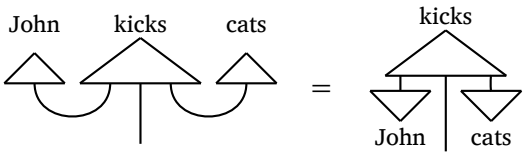
For example, as described in Section 2.1, transitive verbs have type $n^r s n^l$, and can, therefore, be represented in **FHilb** as a rank 3 space $N \otimes S \otimes N$. The transitive sentence *John kicks cats* has type $n(n^r s n^l)n$, which reduces to the sentence type via $\epsilon^r \otimes 1_s \otimes \epsilon^l$. So representing $|\textit{kicks}\rangle$ by:

$$|\textit{kicks}\rangle = \sum_{ijk} c_{ijk} |e_i\rangle \otimes |s_j\rangle \otimes |e_k\rangle$$

using the definitions of the counits in **FHilb** we then have:

$$\begin{aligned} |\textit{John kicks cats}\rangle &= \epsilon_N \otimes 1_S \otimes \epsilon_N(|\textit{John}\rangle \otimes |\textit{kicks}\rangle \otimes |\textit{cats}\rangle) \\ &= \sum_{ijk} c_{ijk} \langle \textit{John}|e_i\rangle \otimes |s_j\rangle \otimes \langle e_k|\textit{cats}\rangle \\ &= \sum_{j} \sum_{ik} c_{ijk} \langle \textit{John}|e_i\rangle \langle e_k|\textit{cats}\rangle |s_j\rangle \end{aligned}$$

Diagrammatically,



The category **FHilb** is actually a †-compact closed category. A †-compact closed category is a compact closed category with an additional *dagger functor* that is an identity-on-objects involution, satisfying natural coherence conditions. In the graphical calculus, the dagger operation "flips diagrams upside-down". In the case of **FHilb**

the dagger sends a linear map to its adjoint, and this allows us to reason about inner products in a general categorical setting, so that meanings of sentences may be compared using the inner product to calculate the cosine distance between vector representations.

The abstract categorical framework we have introduced allows meanings to be interpreted not just in **FHilb**, but in any †-compact closed category. We will exploit this freedom when we move to density matrices. Detailed presentations of the ideas in this section are given in Coecke *et al.* (2010) and Preller and Sadrzadeh (2011) and an introduction to relevant category theory in Coecke and Paquette (2011).

3 DENSITY MATRICES IN CATEGORICAL COMPOSITIONAL DISTRIBUTIONAL SEMANTICS

3.1 *Positive operators and density matrices*

The methods outlined in Section 2 can be applied to the richer setting of density matrices. Density matrices are used in quantum mechanics to express uncertainty about the state of a system. For unit vector $|v\rangle$, the projection operator $|v\rangle\langle v|$ onto the subspace spanned by $|v\rangle$ is called a *pure state*. Pure states can be thought of as giving sharp, unambiguous information. In general, density matrices are given by a convex sum of pure states, describing a probabilistic mixture. States that are not pure are referred to as *mixed states*. Necessary and sufficient conditions for an operator ρ to encode such a mixture are:

- $\forall v \in V. \langle v|\rho|v\rangle \geq 0$,
- ρ is self-adjoint,[1]
- ρ has trace 1.

Operators satisfying the first two axioms are called *positive operators*. The third axiom ensures that the operator represents a convex mixture of pure states. Relaxing this condition gives us different choices for normalization.

[1] As we are dealing with real-valued positive operators, this condition is necessary.

3.2 *Representing words as positive matrices*

Within standard distributional semantics, words are represented as vectors, where the values on specific dimensions correspond to some function of the frequency with which they co-occur with the words represented by the basis vectors. The vector space induced can be modified or reduced using singular value decomposition or other techniques, where the basis vectors no longer have specific meanings. In order to represent words as density matrices, we first observe that each word vector has a corresponding pure matrix:

$$|cat\rangle \mapsto |cat\rangle \langle cat|$$

Words which are more general can be built up by taking sums over pure matrices. We can think of the meaning of the word *pet* as represented by:

$$\begin{aligned}[\![pet]\!] = & p_d \, |dog\rangle \langle dog| + p_c \, |cat\rangle \langle cat| + p_t \, |tarantula\rangle \langle tarantula| + \dots \\ & \text{where} \quad \forall i . p_i \geq 0 \quad \text{and} \quad \sum_i p_i = 1\end{aligned}$$

In general, we consider the meaning of a word w to be given by a collection of unit vectors $\{|w_i\rangle\}_i$, where each $|w_i\rangle$ represents an instance of the concept expressed by the word. Each $|w_i\rangle$ is weighted by $p_i \in [0,1]$, such that $\sum_i p_i = 1$. These describe the meaning of w as a weighted combination of exemplars. Then the density operator:

$$[\![w]\!] = \sum_i p_i \, |w_i\rangle \langle w_i|$$

represents the word w.

This is an extension of the distributional hypothesis. The coefficients p_i may be determined as a function of the frequency with which each word represented by a pure matrix co-occurs with the word represented by $[\![w]\!]$, for example.

3.3 *The CPM construction*

Applying Selinger's CPM construction (Selinger 2007) to **FHilb** produces a new †-compact closed category in which the states are positive operators. This construction has previously been exploited in a linguistic setting in Kartsaklis (2015), Piedeleu *et al.* (2015), and Balkır *et al.* (2016).

Throughout this section $\mathscr{C}$ denotes an arbitrary †-compact closed category.

Definition 1 (Completely positive morphism)**.** *A $\mathscr{C}$-morphism $\varphi : A^* \otimes A \to B^* \otimes B$ is said to be completely positive (Selinger 2007) if there exists $C \in \mathsf{Ob}(\mathscr{C})$ and $k \in \mathscr{C}(C \otimes A, B)$, such that φ can be written in the form:*

$$(k_* \otimes k) \circ (1_{A^*} \otimes \eta_C \otimes 1_A)$$

Identity morphisms are completely positive, and completely positive morphisms are closed under composition in $\mathscr{C}$, leading to the following:

Definition 2. *If $\mathscr{C}$ is a †-compact closed category then* **CPM**$(\mathscr{C})$ *is a category with the same objects as $\mathscr{C}$ and its morphisms are the completely positive morphisms.*

The †-compact structure required for interpreting language in our setting lifts to **CPM**$(\mathscr{C})$:

Theorem 1. **CPM**$(\mathscr{C})$ *is also a †-compact closed category. There is a functor:*

$$\begin{aligned} \mathsf{E} : \mathscr{C} &\to \mathbf{CPM}(\mathscr{C}) \\ k &\mapsto k_* \otimes k \end{aligned}$$

This functor preserves the †-compact closed structure, and is faithful "up to a global phase" (Selinger 2007).

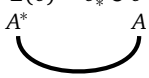
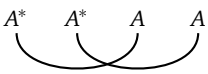
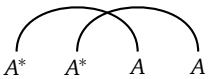
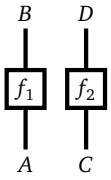
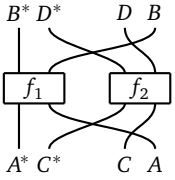
3.4 *Diagrammatic calculus for* **CPM**$(\mathscr{C})$

As **CPM**$(\mathscr{C})$ is also a †-compact closed category, we can use the graphical calculus described in Section 2.3. By convention, the diagrammatic calculus for **CPM**$(\mathscr{C})$ is drawn using thick wires. The corresponding diagrams in $\mathscr{C}$ are given in Table 1.

In the vector space model of meaning the transition between syntax and semantics was achieved by using a strong monoidal functor $\mathsf{Q} : \mathbf{Preg} \to \mathbf{FHilb}$. Language can be assigned semantics in **CPM**(**FHilb**) in an entirely analogous way via a strong monoidal functor:

$$\mathsf{S} : \mathbf{Preg} \to \mathbf{CPM}(\mathbf{FHilb})$$

Table 1: Table of diagrams in **CPM**($\mathscr{C}$) and $\mathscr{C}$

CPM($\mathscr{C}$)	$\mathscr{C}$
$E(\epsilon) = \epsilon_* \otimes \epsilon$ 	$\epsilon : A^* \otimes A^* \otimes A \otimes A \to I$ 
$\epsilon : \lvert e_i\rangle \otimes \lvert e_j\rangle \otimes \lvert e_k\rangle \otimes \lvert e_l\rangle \mapsto \langle e_i \rvert e_k\rangle \langle e_j \rvert e_l\rangle$	
$E(\eta) = \eta_* \otimes \eta$ 	$\eta : I \to A \otimes A \otimes A^* \otimes A^*$ 
$\eta : 1 \mapsto \sum_{ij} \lvert e_i\rangle \otimes \lvert e_j\rangle \otimes \lvert e_i\rangle \otimes \lvert e_j\rangle$	
	
$f_1 \otimes f_2 : A^* \otimes C^* \otimes C \otimes A \to B^* \otimes D^* \otimes D \otimes B$	

Definition 3. *Let $w_1, w_2 \ldots w_n$ be a string of words with corresponding grammatical types t_i in* **Preg**$_{\mathscr{B}}$*. Suppose that the type reduction is given by $t_1, \ldots t_n \xrightarrow{r} x$ for some $x \in \mathsf{Ob}($***Preg**$_{\mathscr{B}})$*. Let $[\![w_i]\!]$ be the meaning of word w_i in* **CPM**(**FHilb**)*, i.e. a state of the form $I \to \mathsf{S}(t_i)$. Then the meaning of $w_1 w_2 \ldots w_n$ is given by:*

$$[\![w_1 w_2 \ldots w_n]\!] = \mathsf{S}(r)([\![w_1]\!] \otimes \ldots \otimes [\![w_n]\!])$$

We now have all the ingredients to derive sentence meanings in **CPM**(**FHilb**).

Example 1. *We firstly show that the results from* **FHilb** *lift to* **CPM**(**FHilb**)*. Let the noun space N be a real Hilbert space with basis vectors given by $\{\lvert n_i\rangle\}_i$, where for some i, $\lvert n_i\rangle = \lvert Clara\rangle$ and for some j, $\lvert n_j\rangle = \lvert beer\rangle$. Let the sentence space be another space S with basis $\{\lvert s_i\rangle\}_i$. The verb $\lvert likes\rangle$ is given by:*

$$\lvert likes\rangle = \sum_{pqr} C_{pqr} \lvert n_p\rangle \otimes \lvert s_q\rangle \otimes \lvert n_r\rangle$$

The density matrices for the nouns Clara and beer are in fact pure states given by:

$$[\![Clara]\!] = |n_i\rangle \langle n_i| \qquad and \qquad [\![beer]\!] = |n_j\rangle \langle n_j|$$

and similarly, $[\![likes]\!]$ *in* **CPM**(**FHilb**) *is:*

$$[\![likes]\!] = \sum_{pqrtuv} C_{pqr} C_{tuv} |n_p\rangle \langle n_t| \otimes |s_q\rangle \langle s_u| \otimes |n_r\rangle \langle n_v|$$

The meaning of the composite sentence is simply $(\varepsilon_N \otimes 1_S \otimes \varepsilon_N)$ *applied to* $([\![Clara]\!] \otimes [\![likes]\!] \otimes [\![beer]\!])$ *as shown in Figure 5, with interpretation in* **FHilb** *shown in Figure 6.*

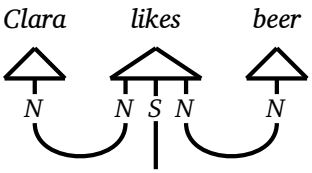


Figure 5:
A transitive sentence in **CPM**($\mathscr{C}$)

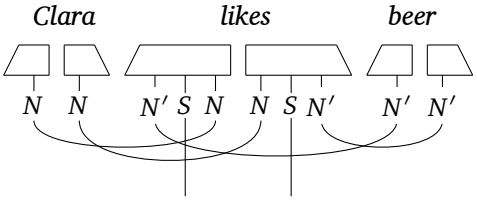


Figure 6:
A transitive sentence in $\mathscr{C}$ with pure states

In terms of linear algebra, this corresponds to:

$$\begin{aligned}[\![Clara\ likes\ beer]\!] &= \varphi([\![Clara]\!] \otimes [\![likes]\!] \otimes [\![beer]\!]) \\ &= \sum_{qu} C_{iqj} C_{iuj} |s_q\rangle \langle s_u|\end{aligned}$$

This is a pure state corresponding to the vector $\sum_q C_{iqj} |s_q\rangle$.

However, in **CPM**(**FHilb**) we can work with more than the pure states.

Example 2. *Let the noun space* N *be a real Hilbert space with basis vectors given by* $\{|n_i\rangle\}_i$. *Let:*

$$|Annie\rangle = \sum_i a_i |n_i\rangle\,,\ |Betty\rangle = \sum_i b_i |n_i\rangle\,,\ |Clara\rangle = \sum_i c_i |n_i\rangle$$

$$|beer\rangle = \sum_i d_i |n_i\rangle, \quad |wine\rangle = \sum_i e_i |n_i\rangle$$

and with the sentence space S, we define:

$$|likes\rangle = \sum_{pqr} C_{pqr} |n_p\rangle \otimes |s_q\rangle \otimes |n_r\rangle$$

$$|appreciates\rangle = \sum_{pqr} D_{pqr} |n_p\rangle \otimes |s_q\rangle \otimes |n_r\rangle$$

Then, we can set:

$$[\![the\ sisters]\!] = \frac{1}{3}(|Annie\rangle\langle Annie| + |Betty\rangle\langle Betty| + |Clara\rangle\langle Clara|)$$

$$[\![drinks]\!] = \frac{1}{2}(|beer\rangle\langle beer| + |wine\rangle\langle wine|)$$

$$[\![enjoy]\!] = \frac{1}{2}(|like\rangle\langle like| + |appreciate\rangle\langle appreciate|)$$

Then, the meaning of the sentence:

$$s = The\ sisters\ enjoy\ drinks$$

is given by:

$$[\![s]\!] = (\varepsilon_N \otimes 1_S \otimes \varepsilon_N)([\![the\ sisters]\!] \otimes [\![enjoy]\!] \otimes [\![drinks]\!])$$

Diagrammatically, this is shown in Figure 7.

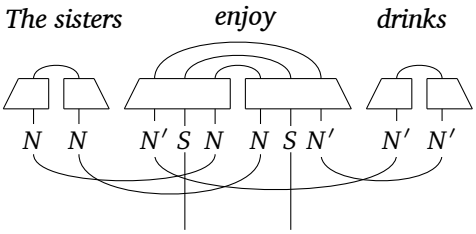


Figure 7: A transitive sentence in $\mathscr{C}$ with impure states

The impurity is indicated by the fact that the pairs of states are connected by wires (Selinger 2007).

4 PREDICATES AND ENTAILMENT

If we consider a model of (non-deterministic) classical computation, a state of a set X is just a subset $\rho \subseteq X$. Similarly, a predicate is a subset $A \subseteq X$. We say that ρ satisfies A if:

$$\rho \subseteq A$$

which we write as $\rho \Vdash A$. Predicate A entails predicate B, written $A \models B$, if for every state ρ:

$$\rho \Vdash A \quad \Rightarrow \quad \rho \Vdash B$$

Clearly this is equivalent to requiring $A \subseteq B$.

4.1 *The Löwner order*

As our linguistic models derive from a quantum mechanical formalism, positive operators form a natural analogue for subsets as our predicates. This follows ideas in D'Hondt and Panangaden (2006) and earlier work in a probabilistic setting in Kozen (1983). Crucially, we can order positive operators (Löwner 1934).

Definition 4 (Löwner order)**.** *For positive operators A and B, we define:*

$$A \sqsubseteq B \iff B - A \textit{ is positive}$$

If we consider this as an entailment relationship, we can follow our intuitions from the non-deterministic setting. Firstly, we introduce a suitable notion of satisfaction. For positive operator A and density matrix ρ, we define $\rho \Vdash A$ as the positive real number $\mathrm{tr}(\rho A)$.

This generalizes satisfaction from a binary relation to a binary function into the positive reals. We then find that the Löwner order can equivalently be phrased in terms of satisfaction as follows:

Lemma 1 (D'Hondt and Panangaden 2006)**.** *Let A and B be positive operators. $A \sqsubseteq B$ if and only if for all density operators ρ:*

$$\rho \Vdash A \quad \leq \quad \rho \Vdash B$$

Linguistically, we can interpret this condition as saying that every noun, for example, satisfies predicate B at least as strongly as it satisfies predicate A.

4.2 *Quantum logic*

Quantum logic (Birkhoff and von Neumann 1936) views the projection operators on a Hilbert space as propositions about a quantum system. As the Löwner order restricts to the usual ordering on projection operators, we can embed quantum logic within the poset of projection operators, providing a direct link to existing theory.

4.3 A general setting for approximate entailment

We can build an entailment preorder on any commutative monoid, viewing the underlying set as a collection of propositions. We then write $A \models B$ and say A entails B if there exists a proposition D such that $A + D = B$. If our commutative monoid is the powerset of some set X, with union the binary operation and unit the empty set, then we recover our non-deterministic computation example from the previous section. If, on the other hand, we take our commutative monoid to be the positive operators on some Hilbert space, with addition of operators and the zero operator as the monoid structure, we recover the Löwner ordering.

In linguistics, we may ask ourselves: does *dog* entail *pet*? Naïvely, the answer is clearly no, not every dog is a pet. This seems too crude for realistic applications though, most dogs are pets, and so we might say *dog* entails *pet* to some extent. This motivates our need for an approximate notion of entailment.

For proposition E, we say that A entails B to the extent E if:

$$A \models B + E$$

We think of E as a error term, for instance in our dogs and pets example, E adds back in dogs that are not pets. Expanding definitions, we find A entails B to extent E if there exists D such that:

$$A + D = B + E \tag{2}$$

From this more symmetrical formulation it is easy to see that for arbitrary propositions A, B, proposition A trivially entails B to extent A, as by commutativity:

$$A + B = B + A$$

It is therefore clear that the mere existence of a suitable error term is not sufficient for a weakened notion of entailment. If we restrict our attention to errors in a complete meet semilattice $\mathscr{E}_{A,B}$, we can take the lower bound on the E satisfying equation (2) as our canonical choice. Finally, if we wish to be able to compare entailment strengths globally, this can be achieved by choosing a partial order $\mathscr{K}$ of "error sizes" and monotone functions:

$$\mathscr{E}_{A,B} \xrightarrow{\kappa_{A,B}} \mathscr{K}$$

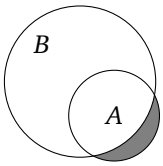
sending errors to their corresponding size.

For example, if A and B are positive operators, we take our complete lattice of error terms $\mathscr{E}_{A,B}$ to be all operators of the form $(1-k)A$ for $k \in [0,1]$, ordered by the size of $1-k$. We then take k as the strength of the entailment, and refer to it as *k-hyponymy*.

In the case of finite sets A, B, we take $\mathscr{E}_{A,B} = \mathscr{P}(A)$, and take the size of the error terms as:

$$\frac{\text{cardinality of } E}{\text{cardinality of } A}$$

measuring "how much" of A we have to supplement B with, as indicated in the shaded region below:



In terms of conditional probability, the error size is then:

$$P(A \mid \neg B)$$

These general error terms are strictly more general than the k-hyponymy.

5 HYPONYMY IN CATEGORICAL COMPOSITIONAL DISTRIBUTIONAL SEMANTICS

Modelling hyponymy in the categorical compositional distributional semantics framework was first considered in Balkır (2014). She introduced an asymmetric similarity measure called *representativeness* on density matrices based on quantum relative entropy. This can be used to translate hyponym-hypernym relations to the level of positive transitive sentences. Our aim here will be to provide an alternative measure which relies only on the properties of density matrices and the fact that they are the states in **CPM**(**FHilb**). This will enable us to quantify the *strength* of the hyponymy relationship, described as k-hyponymy. The measure of hyponymy that we use has an advantage over the representativeness measure. Due to the way it combines with linear maps, we can give a quantitative measure to sentence-level entailment based on the entailment strengths between words, whereas representativeness is not shown to combine in this way.

5.1 *Properties of hyponymy*

Before proceeding with defining the concept of *k-hyponymy*, we give two properties of hyponymy that can be captured by our new measure.

- **Asymmetry.** If A is a hyponym of B, then usually, B is not a hyponym of A.
- **Pseudo-transitivity.** If X is a hyponym of Y and Y is a hyponym of Z, then X is a hyponym of Z. However, if the hyponymy is not perfect, then we get a weakened form of transitivity.

The measure of hyponymy that we described above and named k-hyponymy will be defined in terms of density matrices – the containers for word meanings. The idea is then to define a quantitative order on the density matrices, which is not a partial order, but does give us an indication of the asymmetric relationship between words.

5.2 *Ordering positive matrices*

A density matrix can be used to encode the precision that is needed when describing an action. In the sentence *I took my pet to the vet*, we do not know whether the pet is a dog, cat, tarantula, and so on. The sentence *I took my dog to the vet* is more specific. We then wish to develop an order on density matrices so that *dog*, as represented by $|dog\rangle\langle dog|$ is more specific than *pet* as represented by $[\![pet]\!]$. This ordering may then be viewed as an entailment relation, and entailment between words can lift to the level of sentences, so that the sentence *I took my dog to the vet* entails the sentence *I took my pet to the vet*. Note that we do not require that the sentences have exactly the same structure. For example, we would like *I took my brown dog to the vet* to entail *I took my dog to the vet*, and we would expect this to happen because *brown dog* should entail *dog*.

We now define our notion of approximate entailment, following the discussions of Section 4.3:

Definition 5 (k-hyponym)**.** *We say that A is a k-hyponym of B for a given value of k in the range $(0, 1]$ and write $A \preccurlyeq_k B$ if:*

$$0 \sqsubseteq B - kA$$

Note that such a k need not be unique or even exist at all.

Definition 6 (k_{max} hyponym)**.** *k_{max} is the maximum value of $k \in (0,1]$ for which we have $A \preccurlyeq_{k_{max}} B$.*

In general, we are interested in the maximal value k_{max} for which k-hyponymy holds between two positive operators. This k_{max} value quantifies the strength of the entailment between the two operators.

In what follows, for operator A we write A^+ for the corresponding Moore-Penrose pseudo-inverse and $supp(A)$ for the support of A.

Lemma 2 (Balkır 2014)**.** *Let A, B be positive operators.*

$$supp(A) \subseteq supp(B) \iff \exists k.k > 0 \text{ and } B - kA \geq 0$$

Lemma 3. *For positive self-adjoint matrices A, B such that:*

$$supp(A) \subseteq supp(B)$$

B^+A has non-negative eigenvalues.

We now develop an expression for the optimal k in terms of the matrices A and B.

Theorem 2. *For positive self-adjoint matrices A, B such that:*

$$supp(A) \subseteq supp(B)$$

the maximum k such that $B - kA \geq 0$ is given by $1/\lambda$ where λ is the maximum eigenvalue of B^+A.

Proof. We wish to find the maximum k for which

$$\forall \, |x\rangle \in \mathbb{R}^n . \, \langle x| \, (B - pA) \, |x\rangle \geq 0$$

Since $supp(A) \subseteq supp(B)$, such a k exists. We assume that for $k = 1$, there is at least one $|x\rangle$ such that $\langle x| \, (B - kA) \, |x\rangle \leq 0$, since otherwise we're done. For all $|x\rangle \in \mathbb{R}^n$, $\langle x| \, (B - kA) \, |x\rangle$ increases continuously as k decreases. We therefore decrease k until $\langle x| \, (B - kA) \, |x\rangle \geq 0$, and there will be at least one $|x_0\rangle$ at which $\langle x_0| \, (B - kA) \, |x_0\rangle = 0$. These points are minima so that the vector of partial derivatives $\nabla \, \langle x_0| \, (B - k_0 A) \, |x_0\rangle = 2(B - k_0 A) \, |x_0\rangle = \overrightarrow{0}$ (requires B, A self-adjoint).

Therefore $B \, |x_0\rangle = k_0 A |x_0\rangle$, and so $1/k_0 B^+ B \, |x_0\rangle = B^+ A |x_0\rangle$. Since B^+B is a projector onto the support of B and $supp(A) \subseteq supp(B)$, we have:

$$1/k_0 \, |v_0\rangle = B^+ A \, |v_0\rangle$$

where $|v_0\rangle = B^+ B \, |x_0\rangle$, i.e., $1/k_0$ is an eigenvalue of B^+A.

Now, B^+A has only non-negative eigenvalues, and in fact any pair of eigenvalue $1/k$ and eigenvector $|v\rangle$ will satisfy the condition $B|v\rangle = kA|v\rangle$. We now claim that to satisfy $\forall |x\rangle \in \mathbb{R}^n. \langle x|(B-kA)|x\rangle \geq 0$, we must choose k_0 equal to the reciprocal of the maximum eigenvalue λ_0 of B^+A. For a contradiction, take $\lambda_1 < \lambda_0$, so $1/\lambda_1 = k_1 > k_0 = 1/\lambda_0$. Then we require that $\forall |x\rangle \in \mathbb{R}^n. \langle x|(B-k_1A)|x\rangle \geq 0$, and in particular for $|v_0\rangle$. However:

$$\begin{aligned}\langle v_0|(B-k_1A)|v_0\rangle \geq 0 &\iff \langle v_0|B|v_0\rangle \geq k_1 \langle v_0|A|v_0\rangle \\ &\iff k_0 \langle v_0|A|v_0\rangle \geq k_1 \langle v_0|A|v_0\rangle \\ &\quad \text{contradiction, since } k_0 < k_1\end{aligned}$$

We therefore choose k_0 equal to $1/\lambda_0$ where λ_0 is the maximum eigenvalue of B^+A, and $\langle x|(B-k_0A)|x\rangle \geq 0$ is satisfied for all $|x\rangle \in \mathbb{R}^n$. □

5.3 *Properties of k-hyponymy*

- Reflexivity: k-hyponymy is reflexive for $k = 1$.
- Symmetry: k-hyponymy is neither symmetric nor anti-symmetric.
- Transitivity: k-hyponymy satisfies a version of transitivity. Suppose $A \preccurlyeq_k B$ and $B \preccurlyeq_l C$. Then $A \preccurlyeq_{kl} C$, since:

  $$B \sqsubseteq kA \text{ and } C \sqsubseteq lB \implies C \sqsubseteq klA$$

  by transitivity of the Löwner order.
  For the maximal values k_{max}, l_{max}, m_{max} such that $A \preccurlyeq_{k_{max}} B$, $B \preccurlyeq_{l_{max}} C$ and $A \preccurlyeq_{m_{max}} C$, we have the inequality $m_{max} \geq k_{max}l_{max}$.
- Continuity: For $A \preccurlyeq_k B$, when there is a small perturbation to A, there is a correspondingly small decrease in the value of k. The perturbation must lie in the support of B, but can introduce off-diagonal elements.

Theorem 3. *Given $A \preccurlyeq_k B$ and density operator ρ such that $supp(\rho) \subseteq supp(B)$, then for any $\varepsilon > 0$ we can choose a $\delta > 0$ such that:*

$$A' = A + \delta\rho \implies A' \preccurlyeq_{k'} B \text{ and } |k - k'| < \varepsilon$$

Proof of Theorem 3. We wish to show that we can choose δ such that $|k - k'| < \varepsilon$. We use the notation $\lambda_{max}(A)$ for the maximum eigenvalue of A. $A' = A + \delta\rho$ satisfies the condition of Theorem 2, that

$supp(A') \subseteq supp(B)$, since suppose $|x\rangle \notin supp(B)$. $supp(A) \subseteq supp(B)$, so $|x\rangle \notin supp(A)$ and $A|x\rangle = 0$. Similarly, $\rho|x\rangle = 0$. Therefore $(A+\rho)|x\rangle = A'|x\rangle = 0$, so $|x\rangle \notin supp(A')$.

By Theorem 2 we have:

$$k = \frac{1}{\lambda_{max}(B^+A)}, \quad \text{and} \quad k' = \frac{1}{\lambda_{max}(B^+A')}$$

$$k - k' = \frac{\lambda_{max}(B^+A') - \lambda_{max}(B^+A)}{\lambda_{max}(B^+A')\lambda_{max}(B^+A)} \tag{3}$$

We may treat the denominator of (3) as a constant. We expand the numerator and apply Weyl's inequalities (Weyl 1912). These inequalities apply only to Hermitian matrices, whereas we need to apply these to products of Hermitian matrices. Since B^+, A, and ρ are all real-valued positive semidefinite, the products B^+A and $B^+\rho$ have the same eigenvalues as the Hermitian matrices $A^{\frac{1}{2}}B^+A^{\frac{1}{2}}$ and $\rho^{\frac{1}{2}}B^+\rho^{\frac{1}{2}}$. Now:

$$\begin{aligned}\lambda_{max}(B^+A') - \lambda_{max}(B^+A) &= \lambda_{max}(B^+A + \delta B^+\rho) - \lambda_{max}(B^+A) \\ &\leq \lambda_{max}(B^+A) + \delta\lambda_{max}(B^+\rho) - \lambda_{max}(B^+A) \\ &= \delta\lambda_{max}(B^+\rho) \leq \delta\lambda_{max}(B^+)\lambda_{max}(\rho) \leq \delta\lambda_{max}(B^+)\end{aligned}$$

Therefore:

$$k - k' \leq \delta\frac{\lambda_{max}(B^+)}{\lambda_{max}(B^+A')\lambda_{max}(B^+A)} \tag{4}$$

so that given ε, A, B, we can always choose a δ to make $k - k' \leq \varepsilon$. □

5.4 *Scaling*

When comparing positive operators, in order to standardize the magnitudes resulting from calculations, it is natural to consider normalizing their trace so that we work with density operators. Unfortunately, this is a poor choice when working with the Löwner order as distinct pairs of density operators are never ordered with respect to each other, i.e., for density operators σ, τ, $\sigma \sqsubseteq \tau \Rightarrow \sigma = \tau$. Another option is to bound operators as having maximum eigenvalue 1, as suggested in D'Hondt and Panangaden (2006). With this ordering, the projection operators regain their usual ordering and we recover quantum logic as a suborder of our setting.

Our framework is flexible enough to support other normalization strategies. The optimal choice for linguistic applications is left to future empirical work. Other ideas are also possible. For example we can embed the Bayesian order (Coecke and Martin 2011) within our setting via a suitable transformation on positive operators as follows:

1. Diagonalize the operator, choosing a permutation of the basis vectors such that the diagonal elements are in descending order.
2. Let d_i denote the i^{th} diagonal element. We define the diagonal of a new diagonal matrix inductively as follows:
$$d'_0 = d_0 \qquad\qquad d'_{i+1} = d'_i * d_{i+1}$$
3. Transform the new operator back to the original basis.

Further theoretical investigations of this type are left to future work.

5.5 *Representing the order in the 'Bloch disc'*

The Bloch sphere, Bloch (1946), is a geometrical representation of quantum states. Very briefly, points on the sphere correspond to pure states, and states within the sphere to impure states. Since we consider matrices only over $\mathbb{R}^2$, we disregard the complex phase which allows us to represent the pure states on a circle. A pure state $\cos(\theta/2)|0\rangle + \sin(\theta/2)|1\rangle$ is represented by the vector $(\sin(\theta), \cos(\theta))$ on the circle.

We can calculate the entailment factor k between any two points on the disc. Figure 8 shows contour maps of the entailment strengths for the state with Bloch vector $v = (\frac{3}{4}\sin(\pi/5), \frac{3}{4}\cos(\pi/5))$, using the maximum eigenvalue normalization.

6 RESULTS ON COMPOSITIONALITY

This section provides results and examples on how the notion of hyponymy we have proposed interacts with the compositionality outlined in Section 2. We firstly give an example showing that phrases of different lengths can be compared. We then give a theorem and example to show that our notion of hyponymy 'lifts' to the sentence level, and that the k-values are preserved in a very intuitive fashion.

6.1 *k-hyponymy in phrases of varying length*

We can calculate the extent to which any pair of sentences or phrases are hyponyms of each other. We go back to the simple example in

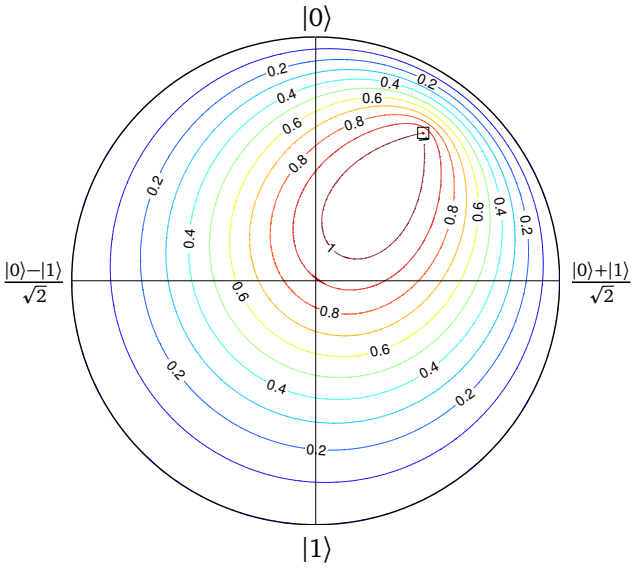


Figure 8:
Entailment strengths
in the Bloch disc for the state
with Bloch vector v

the introduction, comparing 'blond men' to 'men'. Suppose our vector space has basis vectors $|blond\rangle$, $|brunette\rangle$, $|male\rangle$, $|female\rangle$. Then the word 'men' can be given by:

$$[\![men]\!] = \frac{1}{3}(|blond\rangle\langle blond| + |brunette\rangle\langle brunette| + |male\rangle\langle male|)$$

signifying that we are agnostic over all vectors with dimensions $|blond\rangle$, $|brunette\rangle$, $|male\rangle$.

The adjective 'blond' is viewed as an operator which takes nouns to blond nouns. This is given by the following:

$$\begin{aligned}[\![blond_{adj}]\!] = &(|blond\rangle \otimes |blond\rangle)(\langle blond| \otimes \langle blond|) \\ &+ (|blond\rangle \otimes |brunette\rangle)(\langle brunette| \otimes \langle blond|) \\ &+ \sum_{i,j \notin \{blond,brunette\}} (|i\rangle \otimes |i\rangle)(\langle j| \otimes \langle j|)\end{aligned}$$

Then

$$\begin{aligned}[\![blond\ men]\!] &= (1_{N\otimes N} \otimes \epsilon_{N\otimes N})([\![blond_{adj}]\!] \otimes [\![men]\!]) \\ &= \frac{2}{3}|blond\rangle\langle blond| + \frac{1}{3}|male\rangle\langle male|\end{aligned}$$

Then if Carlos is described by the pure state

$$|Carlos\rangle = \frac{1}{\sqrt{2}}(|blond\rangle + |male\rangle)$$

we have

$$[\![Carlos]\!] = |Carlos\rangle \langle Carlos| \preccurlyeq_k [\![blond\ men]\!]$$

for $k = \frac{4}{9}$ by Theorem 2. For Janette described by the pure state $|Janette\rangle = \frac{1}{\sqrt{2}}(|blond\rangle + |female\rangle)$, we have

$$[\![Janette]\!] = |Janette\rangle \langle Janette| \preccurlyeq_k [\![blond\ men]\!]$$

for $k = 0$, since $supp([\![Janette]\!]) \not\subseteq supp([\![blond\ men]\!])$.

An obvious line of enquiry here is to consider how to build this type of adjective operator computationally. One strategy might be to extend the linear regression approach from Baroni and Zamparelli (2010) and Grefenstette *et al.* (2013), having built representations of '*noun*' and the noun phrase '*blond noun*'. Techniques for building density matrix representations of nouns are described in Sadrzadeh *et al.* (2018).

6.2 *Sentence k-hyponymy*

We can show that the application of k-hyponymy to various phrase types holds in the same way. In this section we provide a general proof for varying phrase types. We adopt the following conventions:

- A *positive phrase* is assumed to be a phrase in which individual words are upwardly monotone in the sense described by (Barwise and Cooper 1981; MacCartney and Manning 2007). This means that, for example, the phrase does not contain any negations, including words like *not*.
- The *length* of a phrase is the number of words in it, not counting definite and indefinite articles.

Theorem 4 (Sentence k-hyponymy)**.** *Let Φ and Ψ be two positive phrases of the same length and grammatical structure, expressed in the same noun spaces N and sentence spaces S. Denote the words of Φ, in the order in which they appear, by $A_1, \ldots, A_n$. Similarly, denote these in Ψ by $B_1, \ldots, B_n$. Let their corresponding density matrices be denoted by $[\![A_1]\!], \ldots, [\![A_n]\!]$ and $[\![B_1]\!], \ldots, [\![B_n]\!]$ respectively. Suppose that $[\![A_i]\!] \preccurlyeq_{k_i} [\![B_i]\!]$ for $i \in \{1, \ldots, n\}$ and some $k_i \in (0, 1]$. Finally, let φ be the sentence meaning map for both Φ and Ψ, such that $\varphi(\Phi)$ is the meaning of Φ and $\varphi(\Psi)$ is the meaning of Ψ. Then:*

$$\varphi(\Phi) \preccurlyeq_{k_1 \cdots k_n} \varphi(\Psi)$$

so $k_1 \cdots k_n$ *provides a lower bound on the extent to which* $\varphi(\Phi)$ *entails* $\varphi(\Psi)$.

Proof of Theorem 4. First of all, we have $[\![A_i]\!] \preccurlyeq_{k_i} [\![B_i]\!]$ for $i \in \{1, \ldots, n\}$. This means that for each i, we have positive matrices ρ_i and non-negative reals k_i such that $[\![B_i]\!] = k_i[\![A_i]\!] + \rho_i$. Now consider the meanings of the two sentences. We have:

$$\begin{aligned}
\varphi(\Phi) &= \phi([\![A_1]\!] \otimes \ldots \otimes [\![A_n]\!]) \\
\varphi(\Psi) &= \varphi([\![B_1]\!] \otimes \ldots \otimes [\![B_n]\!]) \\
&= \varphi\left((k_1[\![A_1]\!] + \rho_1) \otimes \ldots \otimes (k_n[\![A_n]\!] + \rho_n\right) \\
&= (k_1 \cdots k_n)\varphi([\![A_1]\!] \otimes \ldots \otimes [\![A_n]\!]) + \varphi(P)
\end{aligned}$$

where P consists of a sum of tensor products of positive matrices, namely:

$$P = \sum_{S \subset \{1,\ldots,n\}} \bigotimes_{i=1}^{n} \sigma_i$$

where:

$$\sigma_i = \begin{cases} k_i[\![A_i]\!] & \text{if } i \in S \\ \rho_i & \text{if } i \notin S \end{cases} \tag{5}$$

Then we have:

$$\varphi(\Psi) - (k_1 \ldots k_n)\varphi(\Phi) = \varphi(P) \geq 0$$

since P is a sum of tensor products of positive matrices, and φ is a completely positive map. Therefore:

$$\varphi(\Phi) \preccurlyeq_{k_1 \cdots k_n} \varphi(\Psi)$$

as required. □

Intuitively, this means that if (some of) the words of a sentence Φ are k-hyponyms of (some of) the words of sentence Ψ, then this hyponymy is translated into sentence hyponymy. Upward-monotonicity is important here, in particular as introduced by some implicit quantifiers. It might be objected that *dogs bark* should not imply *pets bark*. If the implicit quantification is universal, then this is true, however

the universal quantifier is downward monotone in the first argument, and therefore does not conform to the convention concerning positive phrases. If the implicit quantification is existential, then *some dogs bark* does entail *some pets bark*, and the problem is averted. Discussion of the behaviour of quantifiers and other word types is given in, for example, Barwise and Cooper (1981) or MacCartney and Manning (2007).

The quantity $k_1 \cdots k_n$ is not necessarily maximal, and indeed usually is not. As we only have a lower bound, zero entailment strength between a pair of components does not imply zero entailment strength between entire sentences.

Corollary 1. *Consider two sentences:*

$$\Phi = \bigotimes_i [\![A_i]\!] \qquad \Psi = \bigotimes_i [\![B_i]\!]$$

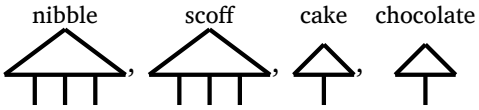
such that for each $i \in \{1, \ldots, n\}$ *we have* $[\![A_i]\!] \sqsubseteq [\![B_i]\!]$*, i.e. there is strict entailment in each component. Then there is strict entailment between the sentences* $\varphi(\Phi)$ *and* $\varphi(\Psi)$*.*

Proof of Corollary 1. Since $k_i = 1$ for each $i = \{1, \ldots, n\}$,

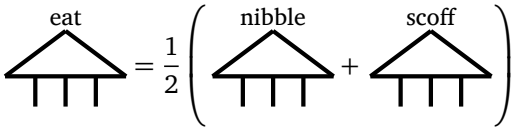
$$\begin{aligned} \varphi(\Phi) \preccurlyeq_{k_1 \cdots k_n} \varphi(\Psi) &\implies \varphi(\Phi) \preccurlyeq_1 \varphi(\Psi) \\ &\implies \varphi(\Phi) \leq \varphi(\Psi) \end{aligned}$$

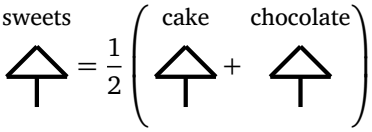
□

We consider a concrete example. Suppose we have a noun space N with basis $\{|e_i\rangle\}_i$, and sentence space S with basis $\{|x_j\rangle\}_j$ We consider the verbs *nibble*, *scoff* and the nouns *cake*, *chocolate*:

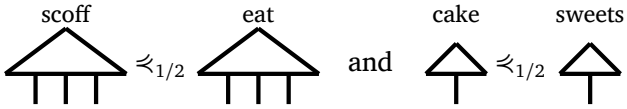


where these nouns and verbs are pure states. The more general *eat* and *sweets* are given by:

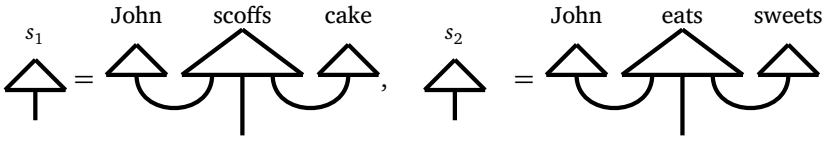




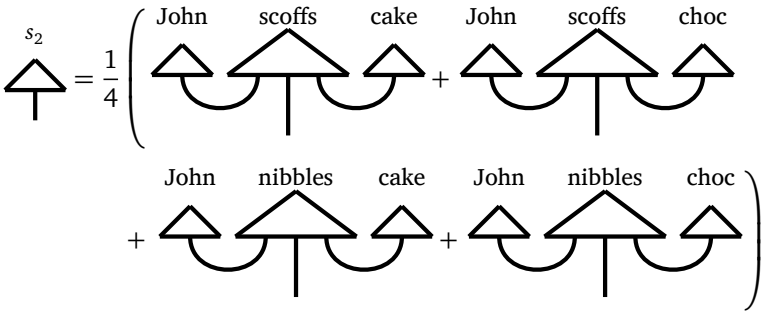
Then



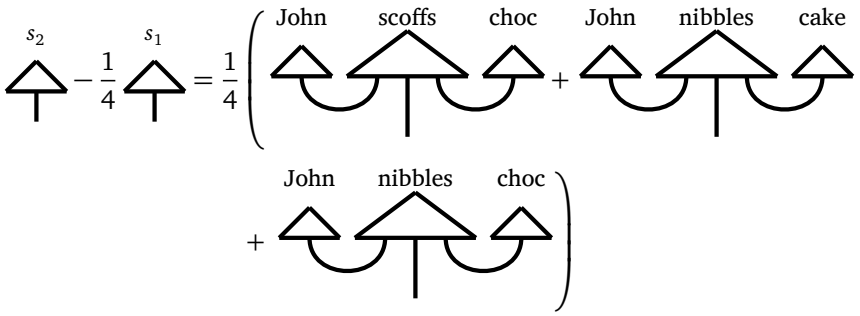
We consider the sentences:



and as per Theorem 4, we will show that $[\![s_1]\!] \preccurlyeq_{kl} [\![s_2]\!]$ where $kl = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. Expanding $[\![s_2]\!]$ we obtain:



Therefore:



We can see that $[\![s_2]\!] - \frac{1}{4}[\![s_1]\!]$ is positive by positivity of the individual elements and the fact that positivity is preserved under addition and tensor product. Therefore $[\![s_1]\!] \preccurlyeq_{kl} [\![s_2]\!]$ as required.

7 A TOY EXPERIMENT

To investigate the effectiveness of the model we perform a toy experiment using a simplified version of the model. We use the dataset introduced in Balkır *et al.* (2016). This dataset consists of pairs of simple sentences annotated by humans as to whether the first sentence entails the second. Example pairs are:

recommend development $\models$ suggest improvement
progress reduce $\models$ development replace

The first sentence is rated highly by humans for entailment, whereas the second has lower ratings. The sentences are either noun-verb or verb-noun, and they are of the same type within the pairs.

We use simplified models of composition which we detail as follows. The first model is a baseline, where we use only the verb to predict the entailment between the two sentences. For the second and third models, we use the notion of a Frobenius algebra. As described in Kartsaklis *et al.* (2012), we can 'lift' lower-order vectors and tensors to higher-order ones. This means that we can obtain a representation for the verb by lifting a density matrix representation. This has the important aspect that the dimensionality needed to represent the word is greatly reduced. In the category **CPM**(**FHilb**), there are two Frobenius algebras we can use. The first equates to a pointwise multiplication of the noun and the verb, and the second is expressed by

$$\rho(s) = \rho(n)^{1/2}\rho(v)\rho(n)^{1/2}$$

where $\rho(s)$, $\rho(n)$, and $\rho(v)$ indicate density matrices for the sentence, noun, and verb respectively.

The last model we examine is an additive model. In general, addition of two positive operators will not be a morphism in **CPM**(**FHilb**). However, in the particular case where the operators are density matrices, we can design a morphism that will implement addition. We give this morphism diagrammatically in Figure 9.

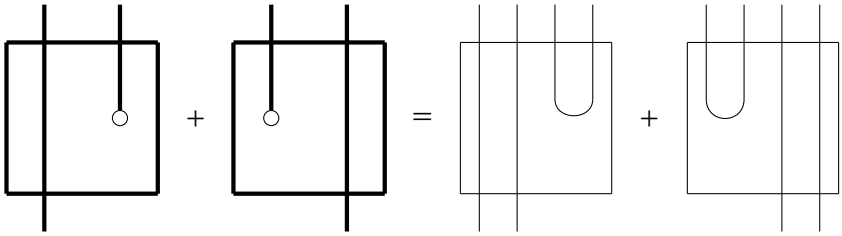


Figure 9: Morphism implementing addition of density matrices

To build density matrices for the nouns and verbs, we firstly collect a set of hyponyms for each word. To do this, we use WordNet (Miller 1995) via the Natural Language ToolKit (nltk) package in Python (Bird *et al.* 2009). We traverse the WordNet graph below each word to a depth of 8, and collect lemma names of every hyponym encountered. We then use GloVe vectors (Pennington *et al.* 2014) to build representations of each word as follows. Firstly, note that in fact the majority of the hyponyms encountered in WordNet were not present in the off-the-shelf GloVe dataset. Approximately 47,000 hyponyms were found across all words in the sentence pairs, of which approximately 10,000 were in the GloVe dataset. To build the density matrix representations for each word, we simply summed the density matrices corresponding to each GloVe vector for each hyponym of the word, and normalised. We added in some small random values along the diagonal, uniformly distributed over $[0, 10^{-3})$ and renormalised. This step is used to ensure that there is some minimal amount of entailment between every word. After creating sentence vectors from the composition of noun and verb vectors, we calculated the entailment using the result from Theorem 2. We ran the experiments over 50, 100, 200, and 300 dimension vectors. We judged the results by computing Spearman's ρ between the generated results and the mean of the human judgements. The best results were obtained with 50 dimensional vectors which we report in Table 2.

Model	ρ	p
Verb-only	0.268	> 0.25
Frobenius mult.	0.508	> 0.05
Frobenius n.c.	0.436	> 0.05
Additive	0.643	> 0.001
Inter-annotator	0.66	–

Table 2:
Results in the sentence entailment task

All the compositional models beat the verb-only baseline. The highest scoring model was the additive model, achieving close to inter-annotator agreement. Note that the sentences were extremely simple, and so it would be good to see how the commutative additive model fares when presented with more complex sentences. The best results from Balkır *et al.* (2016) were $\rho = 0.66$ for a vector-based model using the Spearman's ρ metric and our results are comparable. These vectors were built using part-of-speech information which our model did not use, so there is scope for improvement in that direction.

8 CONCLUSION

Integrating a logical framework with compositional distributional semantics is an important step in improving this model of language. By moving to the setting of density matrices, we have described a graded measure of hyponymy that may be used to describe the extent of hyponymy between two words represented within this enriched framework. This approach extends uniformly to provide hyponymy strengths between two phrases of the same type. That type can be any part of speech for which entailment makes sense, such as a noun phrase, verb phrase, or sentence. This includes pairs of phrases with differing numbers of words. We have also shown how a lower bound on hyponymy strength of phrases of the same structure can be calculated from their components.

Whilst we have given a means for modelling hyponymy in a compositional manner, and provided results on how hyponymy strengths compose, the task of integrating logical and distributional semantics is extremely wide-ranging. We mention here a number of areas to which we can start to contribute.

As mentioned in the introduction, some forms of crisp entailment are based in grammatical structure. So, for example, some adjectives interact with nouns to narrow down concepts, as in our example of 'blond men', and we therefore have that 'blond men' is a hyponym of 'men'. Other adjectives should not operate in this way, such as *former* in *former president*. This phenomenon is related to the notion of downward monotone contexts and the inclusion of negative words like *not*, or negative prefixes. At present, our model cannot effectively account for downward-monotone phenomena. In order to do so, additional

structure, such as some form of involution, must be added to begin to model these phenomena.

The area of grammatical kinds of entailment also includes phenomena such as verb-phrase ellipsis. The framework developed here is all within the category of pregroups, and in order to be able to model more complex grammatical phenomena, we may need to move to other grammar categories. This has started to be developed in Kartsaklis *et al.* (2016) and we may therefore be able to use these methods within our current model.

The area of quantification is an important one. Hedges and Sadrzadeh (2016) have started to develop a theory of quantification within this framework, and so this is an area is which extension could be possible.

Another line of inquiry is to examine transitivity behaves. In some cases entailment can strengthen. We had that *dog* entails *pet* to a certain extent, and that *pet* entails *mammal* to a certain extent, but that *dog* completely entails *mammal*.

Our framework supports different methods of scaling the positive operators representing propositions. Empirical work will be required to establish the most appropriate method in linguistic applications.

ACKNOWLEDGEMENTS

Bob Coecke, Martha Lewis, and Dan Marsden gratefully acknowledge funding from AFOSR grant Algorithmic and Logical Aspects when Composing Meanings. Martha Lewis gratefully acknowledges funding from NWO Veni grant Metaphorical Meanings for Artificial Agents.

REFERENCES

Esma BALKIR (2014), *Using Density Matrices in a Compositional Distributional Model of Meaning*, Master's thesis, University of Oxford, http://www.cs.ox.ac.uk/people/bob.coecke/Esma.pdf.

Esma BALKIR, Mehrnoosh SADRZADEH, and Bob COECKE (2016), Distributional Sentence Entailment Using Density Matrices, in Mohammad T. HAJIAGHAYI and Mohammad R. MOUSAVI, editors, *Topics in Theoretical Computer Science*, volume 9541 of *Lecture Notes in Computer Science*, pp. 1–22, Springer, Cham, https://doi.org/10.1007/978-3-319-28678-5_1.

Dea BANKOVA (2015), *Comparing Meaning in Language and Cognition: P-Hyponymy, Concept Combination, Asymmetric Similarity*, Master's thesis, University of Oxford, http://www.cs.ox.ac.uk/people/bob.coecke/Dea.pdf.

Marco BARONI, Raffaella BERNARDI, Ngoc-Quynh DO, and Chung-chieh SHAN (2012), Entailment above the word level in distributional semantics, in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 23–32, Association for Computational Linguistics, http://aclweb.org/anthology/E12-1004.

Marco BARONI and Roberto ZAMPARELLI (2010), Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space, in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1183–1193, Association for Computational Linguistics, http://aclweb.org/anthology/D10-1115.

Jon BARWISE and Robin COOPER (1981), Generalized Quantifiers and Natural Language, *Linguistics and Philosophy*, 4:159–219.

Steven BIRD, Ewan KLEIN, and Edward LOPER (2009), *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, O'Reilly Media, Inc.

Garrett BIRKHOFF and John VON NEUMANN (1936), The Logic of Quantum Mechanics, *Annals of Mathematics*, 37(4):823–843, ISSN 0003486X, http://www.jstor.org/stable/1968621.

William BLACOE, Elham KASHEFI, and Mirella LAPATA (2013), A Quantum-Theoretic Approach to Distributional Semantics, in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 847–857, Association for Computational Linguistics, http://aclweb.org/anthology/N13-1105.

Felix BLOCH (1946), Nuclear Induction, *Phys. Rev.*, 70:460–474, doi:10.1103/PhysRev.70.460, https://link.aps.org/doi/10.1103/PhysRev.70.460.

Samuel R. BOWMAN, Christopher POTTS, and Christopher D. MANNING (2015), Recursive Neural Networks Can Learn Logical Semantics, in *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pp. 12–21, Association for Computational Linguistics, doi:10.18653/v1/W15-4002, http://aclweb.org/anthology/W15-4002.

Daoud CLARKE (2009), Context-theoretic Semantics for Natural Language: An Overview, in *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, pp. 112–119, Association for Computational Linguistics, Stroudsburg, PA, USA, http://dl.acm.org/citation.cfm?id=1705415.1705430.

Bob Coecke, Edward Grefenstette, and Mehrnoosh Sadrzadeh (2013), Lambek vs. Lambek: Functorial vector space semantics and string diagrams for Lambek calculus, *Annals of Pure and Applied Logic*, 164(11):1079 – 1100, ISSN 0168-0072, https://doi.org/10.1016/j.apal.2013.05.009, special issue on Seventh Workshop on Games for Logic and Programming Languages (GaLoP VII).

Bob Coecke and Keye Martin (2011), A partial order on classical and quantum states, in *New Structures for Physics*, pp. 593–683, Springer.

Bob Coecke and Éric Oliver Paquette (2011), Categories for the practising physicist, in *New Structures for Physics*, pp. 173–286, Springer, https://doi.org/10.1007/978-3-642-12821-9_3.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen J Clark (2010), Mathematical Foundations for a Compositional Distributional Model of Meaning, *Linguistic Analysis*, 36(1):345–384.

Ido Dagan, Oren Glickman, and Bernardo Magnini (2006), The PASCAL Recognising Textual Entailment Challenge, in Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pp. 177–190, Springer, Berlin, Heidelberg, https://doi.org/10.1007/11736790_9.

Ellie D'Hondt and Prakash Panangaden (2006), Quantum Weakest Preconditions, *Mathematical Structures in Computer Science*, 16(3):429–451, https://doi.org/10.1017/S0960129506005251.

Maayan Geffet and Ido Dagan (2005), The Distributional Inclusion Hypotheses and Lexical Entailment, in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 107–114, Association for Computational Linguistics, http://aclweb.org/anthology/P05-1014.

Edward Grefenstette, Georgiana Dinu, Yi Zhang, Mehrnoosh Sadrzadeh, and Marco Baroni (2013), Multi-Step Regression Learning for Compositional Distributional Semantics, in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pp. 131–142, Association for Computational Linguistics, http://aclweb.org/anthology/W13-0112.

Edward Grefenstette and Mehrnoosh Sadrzadeh (2011), Experimental Support for a Categorical Compositional Distributional Model of Meaning, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1394–1404, Association for Computational Linguistics, http://aclweb.org/anthology/D11-1129.

Jules Hedges and Mehrnoosh Sadrzadeh (2016), A Generalised Quantifier Theory of Natural Language in Categorical Compositional Distributional

Semantics with Bialgebras, *CoRR*, abs/1602.01635, http://arxiv.org/abs/1602.01635.

Dimitri KARTSAKLIS (2015), *Compositional Distributional Semantics with Compact Closed Categories and Frobenius Algebras*, Ph.D. thesis, University of Oxford, https://arxiv.org/abs/1505.00138.

Dimitri KARTSAKLIS, Matthew PURVER, and Mehrnoosh SADRZADEH (2016), Verb Phrase Ellipsis using Frobenius Algebras in Categorical Compositional Distributional Semantics, in *DSALT Workshop, European Summer School on Logic, Language and Information*, https://www.eecs.qmul.ac.uk/~mpurver/papers/kartsaklis-et-al16dsalt.pdf.

Dimitri KARTSAKLIS, Mehrnoosh SADRZADEH, and Stephen PULMAN (2012), A Unified Sentence Space for Categorical Distributional-Compositional Semantics: Theory and Experiments, in *Proceedings of COLING 2012: Posters*, pp. 549–558, The COLING 2012 Organizing Committee, http://aclweb.org/anthology/C12-2054.

Graham M. KELLY and Miguel L. LAPLAZA (1980), Coherence for compact closed categories, *Journal of Pure and Applied Algebra*, 19:193 – 213, ISSN 0022-4049, https://doi.org/10.1016/0022-4049(80)90101-2.

Douwe KIELA, Laura RIMELL, Ivan VULIĆ, and Stephen CLARK (2015), Exploiting Image Generality for Lexical Entailment Detection, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 119–124, Association for Computational Linguistics, doi:10.3115/v1/P15-2020, http://aclweb.org/anthology/P15-2020.

Lili KOTLERMAN, Ido DAGAN, Idan SZPEKTOR, and Maayan ZHITOMIRSKY-GEFFET (2010), Directional distributional similarity for lexical inference, *Natural Language Engineering*, 16(4):359—389, https://doi.org/10.1017/S1351324910000124.

Dexter KOZEN (1983), A Probabilistic PDL, in David S. JOHNSON, Ronald FAGIN, Michael L. FREDMAN, David HAREL, Richard M. KARP, Nancy A. LYNCH, Christos H. PAPADIMITRIOU, Ronald L. RIVEST, Walter L. RUZZO, and Joel I. SEIFERAS, editors, *Proceedings of the 15th Annual ACM Symposium on Theory of Computing, 25-27 April, 1983, Boston, Massachusetts, USA*, pp. 291–297, ACM, https://doi.org/10.1145/800061.808758.

Joachim LAMBEK (1997), Type Grammar Revisited, in Alain LECOMTE, François LAMARCHE, and Guy PERRIER, editors, *Logical Aspects of Computational Linguistics, Second International Conference, LACL '97, Nancy, France, September 22-24, 1997, Selected Papers*, volume 1582 of *Lecture Notes in Computer Science*, pp. 1–27, Springer, ISBN 3-540-65751-7, https://doi.org/10.1007/3-540-48975-4_1.

Alessandro LENCI and Giulia BENOTTO (2012), Identifying hypernyms in distributional semantic spaces, in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 75–79, Association for Computational Linguistics, `http://aclweb.org/anthology/S12-1012`.

Karl LÖWNER (1934), Über monotone Matrixfunktionen, *Mathematische Zeitschrift*, 38(1):177–216.

Bill MACCARTNEY and Christopher D. MANNING (2007), Natural Logic for Textual Inference, in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, pp. 193–200, Association for Computational Linguistics, Stroudsburg, PA, USA, `http://dl.acm.org/citation.cfm?id=1654536.1654575`.

George A. MILLER (1995), WordNet: A Lexical Database for English, *Communinications of the ACM*, 38(11):39–41, ISSN 0001-0782, doi:10.1145/219717.219748, `http://doi.acm.org/10.1145/219717.219748`.

Michael A. NIELSEN and Isaac L. CHUANG (2011), *Quantum Computation and Quantum Information: 10th Anniversary Edition*, Cambridge University Press, New York, NY, USA, 10th edition, ISBN 1107002176, 9781107002173.

Jeffrey PENNINGTON, Richard SOCHER, and Christopher MANNING (2014), Glove: Global Vectors for Word Representation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Association for Computational Linguistics, doi:10.3115/v1/D14-1162, `http://aclweb.org/anthology/D14-1162`.

Robin PIEDELEU (2014), *Ambiguity in Categorical Models of Meaning*, Master's thesis, University of Oxford, `http://www.cs.ox.ac.uk/people/bob.coecke/Robin.pdf`.

Robin PIEDELEU, Dimitri KARTSAKLIS, Bob COECKE, and Mehrnoosh SADRZADEH (2015), Open System Categorical Quantum Semantics in Natural Language Processing, in Lawrence S. MOSS and Pawel SOBOCI'NSKI, editors, *6th Conference on Algebra and Coalgebra in Computer Science, CALCO 2015, June 24-26, 2015, Nijmegen, The Netherlands*, volume 35 of *LIPIcs*, pp. 270–289, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, ISBN 978-3-939897-84-2, `https://doi.org/10.4230/LIPIcs.CALCO.2015.270`.

Anne PRELLER and Mehrnoosh SADRZADEH (2011), Bell States and Negative Sentences in the Distributed Model of Meaning, *Electronic Notes in Theoretical Computer Science*, 270(2):141 – 153, ISSN 1571-0661, `https://doi.org/10.1016/j.entcs.2011.01.028`, proceedings of the 6th International Workshop on Quantum Physics and Logic (QPL 2009).

C. J. van RIJSBERGEN (2004), *The Geometry of Information Retrieval*, Cambridge University Press, New York, NY, USA, ISBN 0521838053.

Laura RIMELL (2014), Distributional Lexical Entailment by Topic Coherence, in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 511–519, Association for Computational Linguistics, doi:10.3115/v1/E14-1054, http://aclweb.org/anthology/E14-1054.

Mehrnoosh SADRZADEH, Dimitri KARTSAKLIS, and Esma BALKIR (2018), Sentence entailment in compositional distributional semantics, *Annals of Mathematics and Artificial Intelligence*, 82(4):189–218, https://doi.org/10.1007/s10472-017-9570-x.

Peter SELINGER (2007), Dagger Compact Closed Categories and Completely Positive Maps: (Extended Abstract), *Electronic Notes in Theoretical Computer Science*, 170:139 – 163, ISSN 1571-0661, https://doi.org/10.1016/j.entcs.2006.12.018, proceedings of the 3rd International Workshop on Quantum Programming Languages (QPL 2005).

Julie WEEDS, David WEIR, and Diana MCCARTHY (2004), Characterising Measures of Lexical Distributional Similarity, in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, http://aclweb.org/anthology/C04-1146.

Hermann WEYL (1912), Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung), *Mathematische Annalen*, 71(4):441–479.

Dominic WIDDOWS and Stanley PETERS (2003), Word vectors and quantum logic: Experiments with negation and disjunction, in *Proceedings of Mathematics of Language 8*, pp. 141–154.