A Study of Entanglement in a Categorical Framework of Natural Language

Dimitri Kartsaklis
University of Oxford
Department of Computer Science
Oxford, UK
dimitri.kartsaklis@cs.ox.ac.uk

Mehrnoosh Sadrzadeh
Queen Mary University of London
School of Electronic Engineering and Computer Science
London, UK
mehrnoosh.sadrzadeh@qmul.ac.uk

In both quantum mechanics and corpus linguistics based on vector spaces, the notion of entanglement provides a means for the various subsystems to communicate with each other. In this paper we examine a number of implementations of the categorical framework of Coecke et al. [4] for natural language, from an entanglement perspective. Specifically, our goal is to better understand in what way the level of entanglement of the relational tensors (or the lack of it) affects the compositional structures in practical situations. Our findings reveal that a number of proposals for verb construction lead to almost separable tensors, a fact that considerably simplifies the interactions between the words. We examine the ramifications of this fact, and we show that the use of Frobenius algebras mitigates the potential problems to a great extent. Finally, we briefly examine a machine learning method that creates verb tensors exhibiting a sufficient level of entanglement.

1 Introduction

Category theory in general and compact closed categories in particular provide a high level framework to identify and study universal properties of mathematical and physical structures. Abramsky and Coecke [1], for example, use the latter to provide a structural proof for a class of quantum protocols, essentially recasting the vector space semantics of quantum mechanics in a more abstract way. This and similar kinds of abstraction have made compact closed categories applicable to other fields with vector space semantics, for the case of this paper, corpus linguistics. Here, Coecke et al.[4] used them to unify two seemingly orthogonal semantic models of natural language: a syntax-driven compositional approach as expressed by Lambek [15] and distributional models of meaning based on vector spaces. The latter approach is capable of providing a concrete representation of the meaning of a word, by creating a vector with co-occurrence counts of that word in a corpus of text with all other words in the vocabulary. Distributional models of this form have been proved useful in many natural language processing tasks [23, 17, 16], but in general they do not scale up to larger text constituents such as phrases and sentences. On the other hand, the type-logical approaches to language as introduced in [15] are compositional but unable to provide a convincing model of word meaning.

The unification of the two semantics paradigms is based on the fact that both a type logic expressed as a pregroup [15] and finite dimensional vector spaces share a compact closed structure; so in principle there exists a way to express a grammatical derivation as a morphism that defines mathematical manipulations between vector spaces, resulting in a sentence vector. In [4], the solution was based on a Cartesian product between the pregroup category and the category of finite dimensional vector spaces; later this was recasted in a functorial passage from the former to the latter [19, 3, 10]. The general idea behind any of these frameworks is that the grammatical type of each word determines the vector space where the corresponding vector lives. Words with atomic types, such as nouns, are simple vectors living in N. On the other hand, words with relational types, such as adjectives or verbs, live in tensor product spaces of higher order. For instance, an intransitive verb will be an element of an order-2 space such as

B. Coecke, I. Hasuo & P. Panangaden (Eds.):
Quantum Physics and Logic 2014 (QPL 2014).
EPTCS 172, 2014, pp. 249–261, doi:10.4204/EPTCS.172.17

$N \otimes S$, whereas a transitive verb will live in $N \otimes S \otimes N$. These tensors act on their arguments by *tensor contraction*, a generalization of the familiar notion of matrix multiplication to higher order tensors.

Since every relational word is represented by a tensor, naturally *entanglement* becomes an important issue in these models. Informally speaking, elements of tensor spaces which represent meanings of relational words should be entangled to allow for a so called 'flow of information' (a terminology borrowed from categorical quantum mechanics [1]) among the meanings of words in a phrase or sentence. Otherwise, parts of the meaning of these words become isolated from the rest, leading to unwanted consequences. An example would be that all sentences that have the same verb end up to get the same meaning regardless of the rest of the context, and this is obviously not the case in language. Whereas at least intuitively the above argument makes sense, in some of the language tasks we have been experimenting with, non-entangled tensors have provided very good results. For example, in [8] Grefenstette and Sadrzadeh provide results for verbs that are built from the outer product of their context vectors. These results beat the state of the art of that time (obtained by the same authors in a previous paper [7]) by a considerable difference.

The purpose of the current paper is to provide a preliminary study of the entanglement in corpus linguistics and to offer some explanation why phenomena such as the above have been the case: is this a by-product of the task or the corpus or the specific concrete model? We work with a number of concrete instantiations of the framework in sentence similarity tasks and observe their performances experimentally from an entanglement point of view. Specifically, we investigate a number of models based on the weighted relations method of [7], where a verb matrix is computed as the structural mixing of all subject/object pairs with which it appears in the training corpus. We also test a model trained using linear regression [2]. Our findings for the first case have been surprising. It turns out that, contrary to intuition and despite the fact that the construction method should yield entangled matrices, the results are very close to their rank-1 approximations, that is, they are in effect separable. We further investigate the ramifications of this observation and try to explain the good practical predictions. We then experiment with the linear regression model of [2] and show that the level of entanglement is much higher in the verbs of this model. Finally, we look at a number of Frobenius variations of the weighted relation models, such as the ones presented in [13] and a few new constructions exclusive to this paper. The conclusions here are also surprising, but in a positive way. It seems that Frobenius models are able to overcome the unwanted "no-flow" collapses of the separable verbs by generating a partial flow between the verb and either its subject or its object, depending which dimension they are copying.

2 Quantizing the grammar

The purpose of the categorical framework is to map a grammatical derivation to some appropriate manipulation between vector spaces. In this section we will shortly review how this goal is achieved. Our basic type logic is a *pregroup grammar* [15], built on the basis of a pregroup algebra. This is a partially ordered monoid with unit 1, whose each element p has a left adjoint p^l and a right adjoint p^r. This means that they satisfy the following inequalities:

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

A pregroup grammar is the pregroup freely generated over a set of atomic types, which for this paper will be $\{n, s\}$. Here, type n refers to nouns and noun phrases, and type s to sentences. The atomic types and their adjoints can be combined to create types for *relational words*. The type of an adjective, for example, is $n \cdot n^l$, representing something that inputs a noun (from the right) and outputs another noun. Similarly, the type of a transitive verb $n^r \cdot s \cdot n^l$ reflects the fact that verbs of this kind expect two inputs, one noun at each side. A grammatical reduction then follows from the properties of pregroups and specifically the inequalities in (1) above. The derivation for the sentence 'Happy kids play games' has the following form:

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

We refer to the free pregroup generated by a partially ordered set T as $\mathbf{Preg}_F(T)$. Categorically, this structure conforms to the definition of a non-symmetric *compact closed category*. The inequalities in (1) correspond to the ε and η morphisms of a compact closed category, given as follows:

$$\varepsilon^l : A^l \otimes A \to I \qquad \varepsilon^r : A \otimes A^r \to I \tag{2}$$

$$\eta^l : I \to A \otimes A^l \qquad \eta^r : I \to A^r \otimes A \tag{3}$$

Hence the above grammatical reduction becomes the following morphism:

$$(\varepsilon_n^r \otimes 1_s) \circ (1_n \otimes \varepsilon_n^l \otimes 1_{n^r \cdot s} \otimes \varepsilon_n^l) \tag{4}$$

Category $\mathbf{Preg}_F(T)$ is posetal, which means that there is at most one morphism between two given objects. To make this into a full-blown category we work with the free compact closed category generated over T, as described in [20], which we will denote $\mathbf{C}_F(T)$. Furthermore, let us refer to the category of finite-dimensional vector spaces and linear maps over $\mathbb{R}$ as $\mathbf{FVect}_W$, where W is our basic distributional vector space with an orthonormal basis $\{w_i\}_i$. This category is again compact closed (although a symmetric one, since $W \cong W^*$), with the ε and η maps given as follows:

$$\varepsilon^l = \varepsilon^r : W \otimes W \to \mathbb{R} :: \sum_{ij} c_{ij} (\overrightarrow{w_i} \otimes \overrightarrow{w_j}) \mapsto \sum_{ij} c_{ij} \langle \overrightarrow{w_i} | \overrightarrow{w_j} \rangle \tag{5}$$

$$\eta^l = \eta^r : \mathbb{R} \to W \otimes W :: 1 \mapsto \sum_i \overrightarrow{w_i} \otimes \overrightarrow{w_i} \tag{6}$$

The transition from a pregroup reduction to a morphism between vector spaces is achieved by a *strongly monoidal functor* $\mathscr{F} : \mathbf{C}_F(T) \to \mathbf{FVect}_W$ that preserves the compact structure so that $\mathscr{F}(A^l) = \mathscr{F}(A)^l$ and $\mathscr{F}(A^r) = \mathscr{F}(A)^r$. Further, since $\mathbf{FVect}_W$ is symmetric and W has a fixed basis, we have that $\mathscr{F}(A)^r = \mathscr{F}(A)^l \cong \mathscr{F}(A)$. As motivated in previous work [13], we assume that $\mathscr{F}$ assigns the basic vector space W to both of the atomic types, that is we have:

$$\mathscr{F}(n) = \mathscr{F}(s) = W \tag{7}$$

The partial orders between the atomic types are mapped to linear maps from W to W by functoriality. The adjoints of atomic types are also mapped to W, whereas the complex types are mapped to tensor products of vector spaces:

$$\mathscr{F}(n \cdot n^l) = \mathscr{F}(n^r \cdot s) = W \otimes W \qquad \mathscr{F}(n^r \cdot s \cdot n^l) = W \otimes W \otimes W \tag{8}$$

We are now in position to define the meaning of a sentence $w_1 w_2 \ldots w_n$ with type reduction α as follows:

$$\mathscr{F}(\alpha)(\overrightarrow{w_1} \otimes \overrightarrow{w_2} \otimes \ldots \otimes \overrightarrow{w_n}) \tag{9}$$

For example, the meaning of the sentence 'happy kids play games', which has the grammatical reduction (4), is computed as follows:

$$\mathscr{F}\left[(\varepsilon_n^r \otimes 1_s) \circ (1_n \otimes \varepsilon_n^l \otimes 1_{n^r \cdot s} \otimes \varepsilon_n^l)\right] \left(\overline{happy} \otimes \overrightarrow{kids} \otimes \overline{play} \otimes \overrightarrow{games}\right) =$$

$$(\varepsilon_W \otimes 1_W) \circ (1_W \otimes \varepsilon_W \otimes 1_{W \otimes W} \otimes \varepsilon_W) \left(\overline{happy} \otimes \overrightarrow{kids} \otimes \overline{play} \otimes \overrightarrow{games}\right)$$

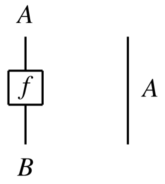
The above categorical computations simplify to the following form:

$$(\overline{happy} \times \overrightarrow{kids})^{\mathsf{T}} \times \overline{play} \times \overrightarrow{games} \tag{10}$$

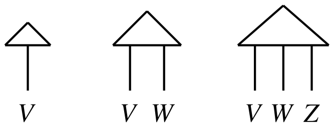
where symbol × denotes tensor contraction and the above is a vector living in our basic vector space W.

3 Pictorial calculus

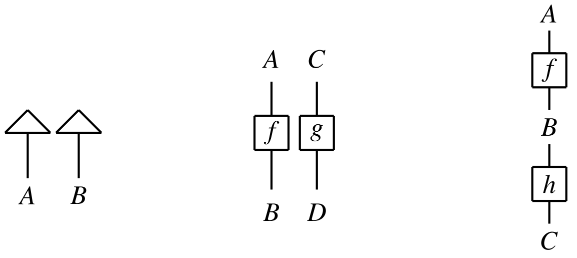
Compact closed categories are complete with regard to a pictorial calculus [14, 24], which can be used for visualizing the derivations and simplifying the computations. We introduce the fragment of calculus that is relevant to the current paper. A morphism $f : A \to B$ is depicted as a box with incoming and outgoing wires representing the objects; the identity morphism $1_A : A \to A$ is a straight line.



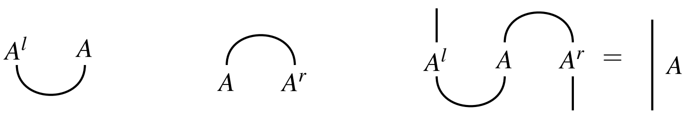
Recall that the objects of $\mathbf{FVect}_W$ are vector spaces. However, for our purposes it is also important to access individual vectors within a vector space. In order to do that, we represent a vector $\overrightarrow{v} \in V$ as a morphism $\overrightarrow{v} : I \to V$. The unit object is depicted as a triangle, while the number of wires emanating from it denotes the order of the corresponding tensor.



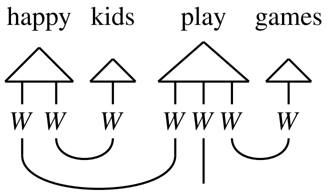
Tensor products of objects and morphisms are depicted by juxtaposing the corresponding diagrams side by side. Composition, on the other hand, is represented as a vertical superposition. For example, from left to right, here are the pictorial representations of the tensor of a vector in A with a vector in B, a tensor of morphisms $f \otimes g : A \otimes C \to B \otimes D$, and a composition of morphisms $h \circ f$ for $f : A \to B$ and $h : B \to C$:



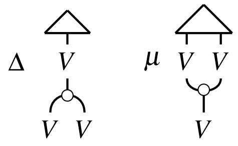
The ε-maps are represented as cups (∪) and the η-maps as caps (∩). Equations such as $(\varepsilon^l_A \otimes 1_{A^r}) \circ (1_{A^l} \otimes \eta^r_A) = 1_A$ now get an intuitive visual justification:



We are now in position to provide a diagram for the meaning of the sentence 'happy kids play games'.



We conclude this section with one more addition to our calculus. As in most quantum protocols, some times the flow of information in linguistics requires elements of classical processing; specifically, we will want the ability to *copy* and *delete* information, which can be provided by introducing *Frobenius algebras*. In **FVect**, any vector space V with a fixed basis $\{\overrightarrow{v_i}\}$ has a Frobenius algebra over it given by Eqs. 11 below.



$$\Delta :: \overrightarrow{v_i} \mapsto \overrightarrow{v_i} \otimes \overrightarrow{v_i} \qquad \iota :: \overrightarrow{v_i} \mapsto 1 \tag{11}$$

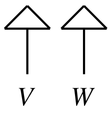
$$\mu :: \overrightarrow{v_i} \otimes \overrightarrow{v_j} \mapsto \delta_{ij}\overrightarrow{v_i} := \begin{cases} \overrightarrow{v_i} & i = j \\ \overrightarrow{0} & i \neq j \end{cases} \qquad \zeta :: 1 \mapsto \sum_i \overrightarrow{v_i}$$

4 Entanglement in quantum mechanics and linguistics

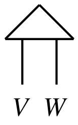
Given two non-interacting quantum systems A and B, where A is in state $|\psi\rangle_A$ and B in state $|\psi\rangle_B$, we denote the state of the composite system $A \otimes B$ by $|\psi\rangle_A \otimes |\psi\rangle_B$. States of this form that can be expressed as the tensor product of two state vectors are called *product* states, and they constitute a special case of separable states. In general, however, the state of a composite system is not necessarily a product state or even a separable one. Fixing bases $\{|i\rangle_A\}$ and $\{|j\rangle_B\}$ for the vector spaces of the two states, a general composite state (separable or not) is denoted as follows:

$$|\psi\rangle_{AB} = \sum_{ij} c_{ij} |i\rangle_A \otimes |j\rangle_B \tag{12}$$

In the case of a pure quantum state, $|\psi\rangle_{AB}$ is separable only if it can be expressed as the tensor product of two vectors; otherwise it is *entangled*. In a similar way, the tensor of a relational word is separable if it is equal to the tensor product of two vectors. In our graphical calculus, these objects are depicted by the juxtaposition of two or more triangles:

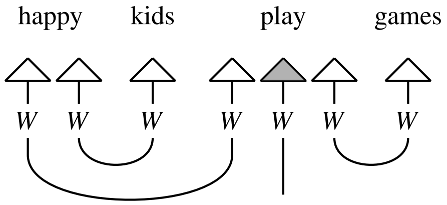


In general, a tensor is not separable if it is a linear combination of many separable tensors. The number of separable tensors needed to express the original tensor is equal to the *tensor rank*. Graphically, a tensor of this form is shown as a single triangle with two or more legs:



5 Consequences of separability

In categorical quantum mechanics terms, entangled states are necessary to allow the flow of information between the different subsystems. In this section we show that the same is true for linguistics. Consider the diagram of our example derivation, where all relational words are now represented by separable tensors (in other words, no entanglement is present):



In this version, the ε-maps are completely detached from the components of the relational tensors that carry the results (left-hand wire of the adjective and middle wire of the verb); as a consequence, flow of information is obstructed, all compositional interactions have been eliminated, and the meaning of the sentence is reduced to the middle component of the verb (shaded vector) multiplied by a scalar, as follows (superscripts denote the left-hand, middle, and right-hand components of separable tensors):

$$\langle\overrightarrow{happy}^{(r)}|\overrightarrow{kids}\rangle\langle\overrightarrow{happy}^{(l)}|\overrightarrow{play}^{(l)}\rangle\langle\overrightarrow{play}^{(r)}|\overrightarrow{games}\rangle\overrightarrow{play}^{(m)}$$

Depending on how one measures the distance between two sentences, this is a very unwelcome effect, to say the least. When using cosine distance, the meaning of all sentences with ‘play’ as the verb will be exactly the same and equal to the middle component of the ‘play’ tensor. For example, the sentence “trembling shadows play hide-and-seek” will have the same meaning as our example sentence. Similarly, the comparison of two arbitrary transitive sentences will be reduced to comparing just the middle components of their verb tensors, completely ignoring any surrounding context. The use of Euclidean distance instead of cosine would slightly improve things, since now we would be at least able to also detect differences in the magnitude between the two middle components. Unfortunately, this metric has been proved not very appropriate for distributional models of meaning, since in the vastness of a highly dimensional space every point ends up to be almost equidistant from all the others. As a result, most implementations of distributional models prefer the more relaxed metric of cosine distance which is length-invariant. Table 1 presents the consequences of separability in a number of grammatical constructs.

Structure	**Simplification**	**Cos-measured result**
adjective-noun	$\overline{adj}\times\overrightarrow{noun}=(\overrightarrow{adj}^{(l)}\otimes\overrightarrow{adj}^{(r)})\times\overrightarrow{noun}=\langle\overrightarrow{adj}^{(r)}\|\overrightarrow{noun}\rangle\cdot\overrightarrow{adj}^{(l)}$	$\overrightarrow{adj}^{(l)}$
intrans. sentence	$\overrightarrow{subj}\times\overline{verb}=\overrightarrow{subj}\times(\overrightarrow{verb}^{(l)}\otimes\overrightarrow{verb}^{(r)})=\langle\overrightarrow{subj}\|\overrightarrow{verb}^{(l)}\rangle\cdot\overrightarrow{verb}^{(r)}$	$\overrightarrow{verb}^{(r)}$
verb-object	$\overline{verb}\times\overrightarrow{obj}=(\overrightarrow{verb}^{(l)}\otimes\overrightarrow{verb}^{(r)})\times\overrightarrow{obj}=\langle\overrightarrow{verb}^{(r)}\|\overrightarrow{obj}\rangle\cdot\overrightarrow{verb}^{(l)}$	$\overrightarrow{verb}^{(l)}$
transitive sentence	$\overrightarrow{subj}\times\overline{verb}\times\overrightarrow{obj}=\overrightarrow{subj}\times(\overrightarrow{verb}^{(l)}\otimes\overrightarrow{verb}^{(m)}\otimes\overrightarrow{verb}^{(r)})\times\overrightarrow{obj}\quad=$ $\langle\overrightarrow{subj}\|\overrightarrow{verb}^{(l)}\rangle\cdot\langle\overrightarrow{verb}^{(r)}\|\overrightarrow{obj}\rangle\cdot\overrightarrow{verb}^{(m)}$	$\overrightarrow{verb}^{(m)}$

Table 1: Consequences of separability in various grammatical structures. Superscripts (l), (m) and (r) refer to left-hand, middle, and right-hand component of a separable tensor

6 Concrete models for verb tensors

Whereas for the vector representations of atomic words of language one can use the much-experimented-with methods of distributional semantics, the tensor representations of relational words is a by-product of the categorical framework whose concrete instantiations are still being investigated. A number of concrete implementations have been proposed so far, e.g. see [7, 13, 9, 12]. These constructions vary from corpus-based methods to machine learning techniques. One problem that researchers have had to address is that tensors of order higher than 2 are difficult to create and manipulate. A transitive verb, for example, is represented by a cuboid living in $W^{\otimes 3}$; if the cardinality of our basic vector space is 1000 (and assuming a standard floating-point representation of 8 bytes per real number), the space required for just a single verb becomes 8 gigabytes. A workaround to this issue is to initially create the verb as a matrix, and then expand it to a tensor of higher order by applying Frobenius Δ operators–that is, leaving one or more dimensions of the resulting tensor empty (filled with zeros).

A simple and intuitive way to create a matrix for a relational word is to structurally mix the arguments with which this word appears in the training corpus [7]. For a transtive verb, this would be given us:

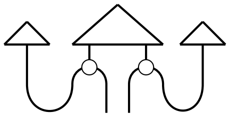
$$\overline{verb} = \sum_i (\overrightarrow{subject_i} \otimes \overrightarrow{object_i}) \tag{13}$$

where $\overrightarrow{subject_i}$ and $\overrightarrow{object_i}$ are the vectors of the subject/object pair for the ith occurrence of the verb in the corpus. The above technique seems to naturally result in an entangled matrix, assuming that the family of subject vectors exhibit a sufficient degree of linear independence, and the same is true for the family of object vectors. Compare this to a straightforward variation which naturally results in a separable matrix, as follows:

$$\overline{verb} = \left(\sum_i \overrightarrow{subject_i}\right) \otimes \left(\sum_i \overrightarrow{object_i}\right) \tag{14}$$

In what follows, we present a number of methods to embed the above verbs from tensors of order 2 to tensors of higher order, as required by the categorical framework.

Relational In [7], the order of a sentence space depends on the arity of the verb of the sentence; for a transitive sentence the result will be a matrix, for an intransitive one it will be a vector, and so on. For the transitive case, the authors expand the original verb matrix to a tensor of order 4 (since now $S = N \otimes N$, the original $N \otimes S \otimes N$ space becomes $N^{\otimes 4}$) by copying both dimensions using Frobenius Δ operators as shown below:

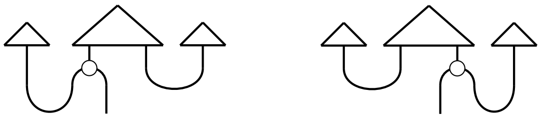


Linear-algebraically, the meaning of a transitive sentence is a matrix living in $W \otimes W$ obtained by the following equation:

$$\overline{subj\ verb\ obj} = (subj \otimes obj) \odot \overline{verb} \tag{15}$$

where the symbol $\odot$ denotes element-wise multiplication.

Frobenius The above method has the limitation that sentences of different structures live in spaces of different tensor orders, so a direct comparison thereof is not possible. As a solution, Kartsaklis et al. [13] propose the copying of only one dimension of the original matrix, which leads to the following two possibilities:



The result is now a vector, computed in the following way, respectively for each case:

$$\textbf{Copy-subject:} \quad \overrightarrow{subj\ verb\ obj} = \overrightarrow{subj} \odot (\overline{verb} \times \overrightarrow{obj}) \tag{16}$$

$$\textbf{Copy-object:} \quad \overrightarrow{subj\ verb\ obj} = \overrightarrow{obj} \odot (\overline{verb}^{\mathsf{T}} \times \overrightarrow{subj}) \tag{17}$$

Each one of the vectors obtained from Eqs. 16 and 17 above addresses a partial interaction of the verb with each argument. It is reasonable then to further combine them in order to get a more complete representation of the verb meaning (and hence the sentence meaning). We therefore define three more models, in which this combination is achieved through vector addition (**Frobenius additive**), element-wise multiplication (**Frobenius multiplicative**), and tensor product (**Frobenius tensored**) of the above.

We conclude this section with two important comments. First, although the use of a matrix for representing a transitive verb might originally seem as a violation of the functorial relation with a pregroup grammar, this is not the case in practice; the functorial relation is restored through the use of the Frobenius operators, which produce a tensor of the correct order, as required by the grammatical type. Furthermore, this notion of “inflation” has the additional advantage that can also work from a reversed perspective: a matrix created by Eq. 13 can be seen as an order-3 tensor originally in $N \otimes S \otimes N$ where the S dimension has been discarded by a ζ Frobenius map. Using this approach, Sadrzadeh and colleagues provide intuitive analyses for wh-movement phenomena and discuss compositional treatments of constructions containing relative pronouns [21, 22].

Finally, we would like to stress out the fact that, despite of the actual level of entanglement in our original verb matrix created by Eq. 13, the use of Frobenius operators as described above equips the inflated verb tensors with an extra level of entanglement in any case. As we will see in Sect. 8 when discussing the results of the experimental work, this detail will be proven very important in practice.

7 Experiments

7.1 Creating a semantic space

Our basic vector space is trained from the ukWaC corpus [5], originally using as a basis the 2,000 content words with the highest frequency (but excluding a list of stop words as well as the 50 most frequent content words since they exhibit low information content). As context we considered a 5-word window from either side of the target word, whereas for our weighting scheme we used local mutual information (i.e. point-wise mutual information multiplied by raw counts). The vector space was normalized and projected onto a 300-dimensional space using singular value decomposition (SVD). These choices are based on our best results in a number of previous experiments [12, 11].

7.2 Detecting sentence similarity

In this section we test the various compositional models of Sect. 6 in two similarity tasks involving pairs of transitive sentences; for each pair, we construct composite vectors for the two sentences, and then we measure their semantic similarity using cosine distance and Euclidean distance. We then evaluate the correlation of each model's performance with human judgements, using Spearman's ρ. In the first task [7], the sentences to be compared are constructed using the same subject and object and semantically correlated verbs, such as 'spell' and 'write'; for example, 'pupils write letters' is compared with 'pupils spell letters'. The dataset consists of 200 sentence pairs.

We are especially interested in measuring the level of entanglement in our verb matrices as these are created by Eq. 13. In order to achieve that, we compute the *rank-1 approximation* of all verbs in our dataset. Given a verb matrix $\overline{verb}$, we first compute its SVD so that $\overline{verb} = \mathbf{U}\Sigma\mathbf{V}^{\mathsf{T}}$, and then we approximate this matrix by using only the highest eigenvalue and the related left and right singular vectors, so that $\overline{verb}_{R1} = \mathbf{U}_1\Sigma_1\mathbf{V}_1^{\mathsf{T}}$. We compare the composite vectors created by the original matrix (Eq. 13), their rank-1 approximations, and the results of the separable model of Eq. 14. We also use a number of baselines: in the 'verbs-only' model, we compare only the verbs (without composing them with the context), while in the additive and multiplicative models we construct the sentence vectors by simply adding and element-wise multiplying the distributional vectors of their words.

The results (Table 2) revealed a striking similarity in the performances of the entangled and separable versions. Using cosine distance, all three models (relational, rank-1 approximation, separable model) have essentially the same behaviour; with Euclidean distance, the relational model performs again the same as its rank-1 approximation, while this time the separable model is lower.

Model	**ρ with cos**	**ρ with Eucl.**
Verbs only	0.329	0.138
Additive	0.234	0.142
Multiplicative	0.095	0.024
Relational	0.400	0.149
Rank-1 approx. of relational	0.402	0.149
Separable	0.401	0.090
Copy-subject	0.379	0.115
Copy-object	0.381	0.094
Frobenius additive	**0.405**	0.125
Frobenius multiplicative	0.338	0.034
Frobenius tensored	**0.415**	0.010
Human agreement	0.60	

Table 2: Results for the first dataset (same subjects/objects, semantically related verbs)

.

The inevitable conclusion that Eq. 13 actually produces a separable matrix was further confirmed by an additional experiment: we calculated the average cosine similarity of the original matrices with their rank-1 approximations, a computation that revealed a similarity as high as 0.99. Since this result might obviously depend on the form of the noun vectors used for creating the verb matrix, this last experiment was repeated with a number of variations of our basic vector space, getting in every case similarities between verb matrices and their rank-1 approximations higher than 0.97. The observed behaviour can only be explained with the presence of a very high level of linear dependence between the subject vectors and between the object vectors. If every subject vector can be expressed as a linear combination of a small number of other vectors (and the same is true for the family of object vectors), then this would drastically reduce the entanglement of the matrix to the level that it is in effect separable.

Our observations are also confirmed in the second sentence similarity task. Here, we use a variation of one of the datasets in [12], consisting of 108 pairs of transitive sentences. The difference with our first task is that now the sentences of a pair are unrelated in a word level, i.e. subjects, objects, and verbs are all different. The results for this second experiment are presented in Table 3.

Model	ρ with cos	ρ with Eucl.
Verbs only	0.449	0.392
Additive	0.581	0.542
Multiplicative	0.287	0.109
Relational	0.334	0.173
Rank-1 approx. of relational	0.333	0.175
Separable	0.332	0.105
Copy-subject	0.427	0.096
Copy-object	0.198	0.144
Frobenius additive	0.428	0.117
Frobenius multiplicative	0.302	0.041
Frobenius tensored	0.332	0.042
Human agreement	0.66	

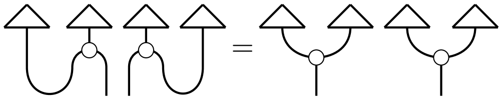
Table 3: Results for the second dataset (different subjects, objects and verbs)

.

As a general observation about the performance of the various models in the two tasks, we note the high scores achieved by the Frobenius models when one uses the preferred method of measurement, that of cosine similarity. Especially the **Frobenius additive** has been proved to perform better than the Relational model, having the additional advantage that it allows comparison between sentences of different structures (since every sentence vector lives in W).

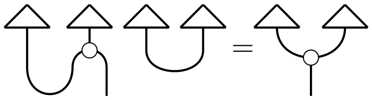
8 Discussion

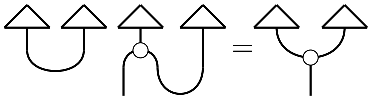
The experiments of Sect. 7 revealed an unwelcome property of a method our colleagues and we have used in the past for creating verb tensors in the context of compositional models [7, 13, 12]. The fact that the verb matrix is in effect separable introduces a number of simplifications in the models presented in Sect. 6. More specifically, the Relational model of [7] is reduced to the following:



$$\overline{subj\ verb\ obj} = (\overrightarrow{subj} \odot \overrightarrow{verb}^{(l)}) \otimes (\overrightarrow{verb}^{(r)} \odot \overrightarrow{obj})$$

Furthermore, the Frobenius models of [13] get these forms:





which means, for example, that the actual equation behind the successful Frobenius additive model is

$$\overrightarrow{subj\ verb\ obj} = (\overrightarrow{subj} \odot \overrightarrow{verb}^{(l)}) + (\overrightarrow{verb}^{(r)} \odot \overrightarrow{obj}) \tag{18}$$

Despite the simplifications presented above, note that none of these models degenerates to the level of producing "constant" vectors or matrices, as argued for in Sect. 5. Indeed, especially in the first task (Table 2) the Frobenius models present top performance, and the relational models follow closely. The reason behind this lies in the use of Frobenius Δ operators for copying the original dimensions of the verb matrix, a computation that equipped the fragmented system with flow, although not in the originally intended sense. The compositional structure is still fragmented into two parts, but at least now the copied dimensions provide a means to deliver the results of the two individual computations that take place, one for the left-hand part of the sentence and one for the right-hand part. Let us see what happens when we use cosine distance in order to compare the matrices of two transitive sentences created with the **Relational** model (the separable version of a verb matrix $\overline{verb}$ is denoted by $\overrightarrow{verb}^{(l)} \otimes \overrightarrow{verb}^{(r)}$):

$$\begin{aligned} \left\langle \overline{subj_1\ verb_1\ obj_1} \middle| \overline{subj_2\ verb_2\ obj_2} \right\rangle &= \\ \left\langle (\overrightarrow{subj}_1 \odot \overrightarrow{verb}_1^{(l)}) \otimes (\overrightarrow{verb}_1^{(r)} \odot \overrightarrow{obj}_1) \middle| (\overrightarrow{subj}_2 \odot \overrightarrow{verb}_2^{(l)}) \otimes (\overrightarrow{verb}_2^{(r)} \odot \overrightarrow{obj}_2) \right\rangle &= \\ \left\langle \overrightarrow{subj}_1 \odot \overrightarrow{verb}_1^{(l)} \middle| \overrightarrow{subj}_2 \odot \overrightarrow{verb}_2^{(l)} \right\rangle \left\langle \overrightarrow{verb}_1^{(r)} \odot \overrightarrow{obj}_1 \middle| \overrightarrow{verb}_2^{(r)} \odot \overrightarrow{obj}_2 \right\rangle & \end{aligned}$$

As also computed and pointed out in [6], the two sentences are broken up to a left-hand part and a right-hand part, and two distinct comparisons take place. As long as we compare sentences of the same structure, as we did here, this method is viable. On the other hand, the **Frobenius** models and their simplifications such as the one in (18) do not have this restriction; in principle, all sentences are represented by vectors living in the same space, so any kind of comparison is possible. In case, however, we do compare sentences of the same structure, these models have the additional advantage that also allow comparisons between *different* sentence parts; this can be seen in the dot product of two sentences created by Eq. 18, which gets the following form:

$$\begin{aligned} \left\langle \overrightarrow{subj}_1 \odot \overrightarrow{verb}_1^{(l)} \middle| \overrightarrow{subj}_2 \odot \overrightarrow{verb}_2^{(l)} \right\rangle + \left\langle \overrightarrow{subj}_1 \odot \overrightarrow{verb}_1^{(l)} \middle| \overrightarrow{verb}_2^{(r)} \odot \overrightarrow{obj}_2 \right\rangle & + \\ \left\langle \overrightarrow{verb}_1^{(r)} \odot \overrightarrow{obj}_1 \middle| \overrightarrow{subj}_2 \odot \overrightarrow{verb}_2^{(l)} \right\rangle + \left\langle \overrightarrow{verb}_1^{(r)} \odot \overrightarrow{obj}_1 \middle| \overrightarrow{verb}_2^{(r)} \odot \overrightarrow{obj}_2 \right\rangle & \end{aligned}$$

9 Using linear regression for entanglement

Corpus-based methods for creating tensors of relational words, such as the models presented so far in this paper, are intuitive and easy to implement. As our experimental work shows, however, this convenience comes with a price. In practice, one would expect that more robust machine learning techniques would produce more reliable tensor representations for composition.

In this section we apply linear regression (following [2]) in order to train verb matrices for a variation of our second experiment, in which we compare elementary verb phrases of the form *verb-object* [18] (so the subjects are dropped). In order to create a matrix for, say, the verb 'play', we first collect all instances of the verb occurring with some object in the training corpus, and then we create non-compositional holistic vectors for these elementary verb phrases following exactly the same methodology as if they were words. We now have a dataset with instances of the form $\langle \overrightarrow{obj_i}, \overrightarrow{play\ obj_i} \rangle$ (e.g. the vector of 'flute' paired with the holistic vector of 'play flute', and so on), that can be used to train a linear regression model in order to produce an appropriate matrix for verb 'play'. The premise of a model like this is that the multiplication of the verb matrix with the vector of a new object will produce a result that approximates the distributional behaviour of all these elementary two-word exemplars used in training. For a given verb, this is achieved by using *gradient descent* in order to minimize the total error between the observed vectors and the vectors predicted by the model, expressed by the following quantity:

$$\frac{1}{2m}\left(\sum_i \overline{verb} \times \overrightarrow{object_i} - \overrightarrow{verb\ object_i}\right)^2 \tag{19}$$

where m is the number of training instances. The average cosine similarity between the matrices we got from this method and their rank-1 approximation was only 0.48, showing that in general the level of entanglement produced by this method is reasonably high. This is also confirmed by the results in Table 4; the rank-1 approximation model presents the worst performance, since, as you might recall from the discussion in Sect. 5, separability here means that every verb-object composition is reduced to the left component of the verb matrix, completely ignoring the interaction with the object.

Model	**ρ with cos**	**ρ with Eucl.**
Verbs only	0.331	0.267
Holistic verb-phrase vectors	0.403	0.214
Additive	0.379	0.385
Multiplicative	0.301	0.229
Linear regression	0.349	0.144
Rank-1 approximation of LR matrices	0.119	0.082
Human agreement	0.55	

Table 4: Results for the verb-phrase similarity task

10 Conclusion

The current study takes a closer look to an aspect of tensor-based compositional models of meaning that so far had escaped the attention of researchers. The discovery that a number of concrete instantiations of the categorical framework proposed in [4] produce relational tensors that are in effect separable stresses the necessity of similar tests for any linear model that follows the same philosophy. Another contribution of this work was that it showed this is not necessarily a bad thing. The involvement of Frobenius operators in the creation of verb tensors equips the compositional structure with the necessary flow, so that a comparison between two sentence vectors can be still carried out between individual parts of each sentence. Therefore, approaches such as the Frobenius additive model proposed in this paper can be still considered as viable and "easy" alternatives to more robust machine learning techniques, such as the gradient optimization technique discussed in Sect. 9.

References

[1] Samson Abramsky & Bob Coecke (2004): *A Categorical Semantics of Quantum Protocols*. In: *19th Annual IEEE Symposium on Logic in Computer Science*, pp. 415–425, doi:10.1109/LICS.2004.1319636.

[2] M. Baroni & R. Zamparelli (2010): *Nouns are Vectors, Adjectives are Matrices*. In: *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[3] B. Coecke, E. Grefenstette & M. Sadrzadeh (2013): *Lambek vs. Lambek: Functorial Vector Space Semantics and String Diagrams for Lambek Calculus*. *Annals of Pure and Applied Logic*, doi:10.1016/j.apal.2013.05.009. Available at http://arxiv.org/abs/1302.0393.

[4] B. Coecke, M. Sadrzadeh & S. Clark (2010): *Mathematical Foundations for Distributed Compositional Model of Meaning. Lambek Festschrift*. *Linguistic Analysis* 36, pp. 345–384.

[5] Adriano Ferraresi, Eros Zanchetta, Marco Baroni & Silvia Bernardini (2008): *Introducing and evaluating ukWaC, a very large web-derived corpus of English*. In: *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pp. 47–54.

[6] E. Grefenstette & M. Sadrzadeh: *Concrete Models and Empirical Evaluations for the Categorical Compositional Distributional Model of Meaning*. *Computational Linguistics*. To appear.

[7] E. Grefenstette & M. Sadrzadeh (2011): *Experimental Support for a Categorical Compositional Distributional Model of Meaning*. In: *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[8] E. Grefenstette & M. Sadrzadeh (2011): *Experimenting with Transitive Verbs in a DisCoCat*. In: *Proceedings of the 2011 EMNLP Workshop on Geometric Models of Natural Language Semantics*.

[9] Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh & Marco Baroni (2013): *Multi-Step Regression Learning for Compositional Distributional Semantics*. In: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*. Available at `http://arxiv.org/abs/1301.6939`.

[10] D. Kartsaklis, M. Sadrzadeh, S. Pulman & B. Coecke (2014): *Reasoning about Meaning in Natural Language with Compact Closed Categories and Frobenius Algebras*. In J. Chubb, A. Eskandarian & V. Harizanov, editors: *Logic and Algebraic Structures in Quantum Computing and Information*, Association for Symbolic Logic Lecture Notes in Logic, Cambridge University Press. To appear.

[11] Dimitri Kartsaklis, Nal Kalchbrenner & Mehrnoosh Sadrzadeh (2014): *Resolving Lexical Ambiguity in Tensor Regression Models of Meaning*. In: *Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics (ACL): Short Papers*, Baltimore, USA.

[12] Dimitri Kartsaklis & Mehrnoosh Sadrzadeh (2013): *Prior Disambiguation of Word Tensors for Constructing Sentence Vectors*. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNL)*, Seattle, USA.

[13] Dimitri Kartsaklis, Mehrnoosh Sadrzadeh & Stephen Pulman (2012): *A Unified Sentence Space for Categorical Distributional-Compositional Semantics: Theory and Experiments*. In: *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012): Posters*, The COLING 2012 Organizing Committee, Mumbai, India, pp. 549–558.

[14] G Maxwell Kelly (1972): *Many-Variable Functorial Calculus (I)*. In G.M. Kelly, M. Laplaza, G. Lewis & S. MacLane, editors: *Coherence in categories*, Springer, pp. 66–105, doi:10.1007/BFb0059556.

[15] J. Lambek (2008): *From Word to Sentence*. Polimetrica, Milan.

[16] T. Landauer & S. Dumais (1997): *A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquision, Induction, and Representation of Knowledge*. *Psychological Review*, doi:10.1037/0033-295X.104.2.211.

[17] C.D. Manning, P. Raghavan & H. Schütze (2008): *Introduction to Information Retrieval*. Cambridge University Press, doi:10.1017/CBO9780511809071.

[18] Jeff Mitchell & Mirella Lapata (2010): *Composition in Distributional Models of Semantics*. *Cognitive Science* 34(8), pp. 1388–1439, doi:10.1111/j.1551-6709.2010.01106.x.

[19] A. Preller & M. Sadrzadeh (2010): *Bell States and Negative Sentences in the Distributed Model of Meaning*. In P. Selinger B. Coecke, P. Panangaden, editor: *Electronic Notes in Theoretical Computer Science, Proceedings of the 6th QPL Workshop on Quantum Physics and Logic*, University of Oxford, doi:10.1016/j.entcs.2011.01.028.

[20] Anne Preller & Joachim Lambek (2007): *Free compact 2-categories*. *Mathematical Structures in Computer Science* 17(2), pp. 309–340, doi:10.1017/S0960129506005901.

[21] Mehrnoosh Sadrzadeh, Stephen Clark & Bob Coecke (2013): *The Frobenius Anatomy of Word Meanings I: Subject and Object Relative Pronouns*. *Journal of Logic and Computation* 23(6), pp. 1293–1317, doi:10.1093/logcom/ext044.

[22] Mehrnoosh Sadrzadeh, Stephen Clark & Bob Coecke (2014): *The Frobenius Anatomy of Word Meanings II: Possessive Relative Pronouns*. *Journal of Logic and Computation*, doi:10.1093/logcom/exu027.

[23] H. Schütze (1998): *Automatic Word Sense Discrimination*. *Computational Linguistics* 24, pp. 97–123.

[24] Peter Selinger (2011): *A Survey of Graphical Languages for Monoidal Categories*. In Bob Coecke, editor: *New structures for physics*, Springer, pp. 289–355.