A Corpus-based Toy Model for DisCoCat

Stefano Gogioso
Quantum Group
University of Oxford
stefano.gogioso@cs.ox.ac.uk

The categorical compositional distributional (DisCoCat) model of meaning rigorously connects distributional semantics and pregroup grammars, and has found a variety of applications in computational linguistics. From a more abstract standpoint, the DisCoCat paradigm predicates the construction of a mapping from syntax to categorical semantics. In this work we present a concrete construction of one such mapping, from a toy model of syntax for corpora annotated with constituent structure trees, to categorical semantics taking place in a category of free R-semimodules over an involutive commutative semiring R.

1 Introduction

The paradigm of distributional semantics draws its roots in the *distributional hypothesis* of [6], evocatively captured by the following words of John Rupert Firth [5]:

You shall know a word by the company it keeps.

In modern computational linguistics, distributional semantics is characterised by the application of statistical methods to large corpora of data. While suitable for certain classes of words (such as nouns), a vanilla statistical approach fails to capture the compositional aspects of language, such as those involving words acting as modifiers (e.g. adjectives and determiners), or words mediating interactions (e.g. verbs and prepositions).

The compositional aspect alone is well covered by the pregroup approach to syntax and grammar of Lambek [10, 11]. The categorical compositional distributional (DisCoCat) model of meaning [4] connects distributional semantics and pregroup grammars, exploiting a close connection between the latter and compact-closed symmetric monoidal categories. The categorical formalism allows for the transfer of tools and diagrammatics from categorical quantum mechanics [1].

Frobenius algebras are key structures in categorical quantum mechanics, where they define orthonormal bases and observables. Through the connection established by the DisCoCat framework, they have also found a variety of applications in computational linguistics. They have been used to model, amongst other things, adjectives and verbs [8], relative pronouns [13, 14] and intonational / informational structure [7]. More generally, they provide an efficient way to represent and manipulate operators on an M-dimensional vector space in time/space linear in M (rather than quadratic). The extension of categorical quantum mechanics from the pure-state case to the mixed-state case via the CPM construction [15] has also found application in linguistics: in recent work, density matrices have been used to model ambiguity [12] and sentence entailment [2].

In this work we construct an abstract categorical model for DisCoCat starting from a generic corpus[1] annotated with constituent structure trees. Concretely, we will work with context-free grammars à la

[1] From our abstract standpoint, semantics depend entirely one the given corpus, no matter the size or quality of annotations. Larger, better annotated corpora will result in better semantics, smaller or poorly annotated corpora will result in worse semantics.

Kartsaklis, D., Lewis, M. and Rimell, L. (Eds):
2016 Workshop on Semantic Spaces at the Intersection of
NLP, Physics and Cognitive Science (SLPCS'16)
EPTCS 221, 2016, pp. 20–28, doi:10.4204/EPTCS.221.3

© S. Gogioso
This work is licensed under the Creative Commons
Attribution-Noncommercial-Share Alike License.

Chomsky [3], but Combinatory Categorial Grammar (CCG) [16] and dependency grammars [9] could also be used. In Section 2 we construct a toy model of syntax for the corpus, as a simplifying intermediate step between constituent structure trees and the categorical semantics. In our toy model of syntax, words are classified according to three possible functions:

(i) *object words*: the basic building blocks (modelled as vectors in the semantic space);

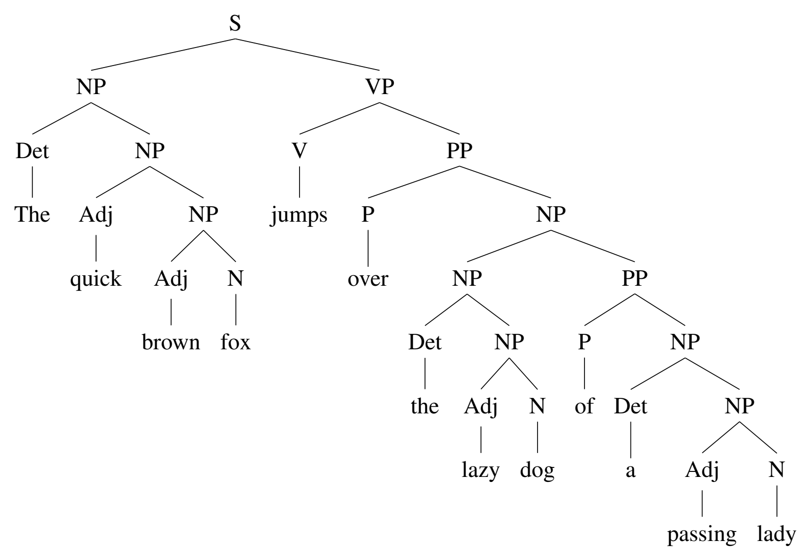
(ii) *modifier words*: modify individual object words (modelled as unary operators on the space);

(iii) *interaction words*: connect distinct object words (modelled as binary operators on the space).

In Section 3, we consider the compact-closed symmetric monoidal category R-Mod of R-semimodules over some involutive commutative semiring R. We model object words as vectors in a free R-semimodule $\mathscr{H}$, which we construct from our intermediate toy model of syntax. Using Frobenius algebras, we model modifier words as unary operators on $\mathscr{H}$, and interaction words as binary operators on $\mathscr{H}$. In Section 4, we discuss some possible future extensions of this model.

2 The toy model of syntax

We begin by considering an abstract annotated corpus. We model this as a set $\mathscr{C}$ of independent sentences, each sentence $\underline{s}$ coming with an annotated tree $T(\underline{s})$ describing its grammatical structure. Concretely, we will have in mind the constituent structure trees from context-free grammars à la Chomsky, as in the following example:



(2.1)

We assume the internal nodes to be labelled from a finite set of **phrasal categories**, and the leaf nodes (corresponding to the words of the sentence) to be labelled from a finite set of **lexical categories**; we will refer to the set of all possible labels, both for internal nodes and for leaf nodes, as the set of **syntactic categories**. Concretely, we will have the following categories in mind:

- **Lexical categories:** Adjectives (Adj), Prepositions (P), Adverbs (Adv), Conjunctions (Conj), Determiners (Det), Nouns (N), Personal Pronouns (Pron), Possessive Pronouns (Poss), Verbs (V).
- **Phrasal categories:** Adjective Phrases (AdjP), Adverb Phrases (AdvP), Adposition Phrases (PP), Noun Phrases (NP), Verb Phrases (VP), Sentences (S).

This set of categories has been chosen for simplicity of exposition: realistic applications would use a richer set (e.g. Penn Treebank tagset). With minor modifications, constituency grammar could be substituted with Combinatory Categorial Grammar (CCG)[2]; similarly, dependency grammars could be used.

We assume the constituent structure trees to be binary, with the exception of nodes involving conjunctions (which we assume to be ternary with sub-trees on the left / right and a leaf of lexical category Conj in the middle). A proper treatment of conjunctions (both coordinating and subordinating) does not appear here, and is left to future work. Leaf nodes will be labelled from the set $\mathrm{Cat_L}$ of lexical categories, and internal nodes will be labelled from the set $\mathrm{Cat_P}$ of phrasal categories. If we denote by W the set of all words appearing in the corpus, each sentence $\underline{s}$ (say $N_{\underline{s}}$ words long) will be modelled by a function $\underline{s} : \{1,...,N_{\underline{s}}\} \to W \times \mathrm{Cat_L}$; because the corpus is annotated, there is no ambiguity in lexical categories for the words in a given sentence. When talking about words, we will henceforth mean pairs $\overline{w} := (w,c) \in W \times \mathrm{Cat_L}$ of a (bare) word tagged with a lexical category. We will denote individual instances of words in sentences of the corpus by $(\underline{s}, j)$, meaning the j-th word instance of sentence $\underline{s} \in \mathscr{C}$.
We will distinguish between three possible functions[3] for words in a sentence:

(i) the **object words**, the ones which will be modelled as vectors in the semantic space;

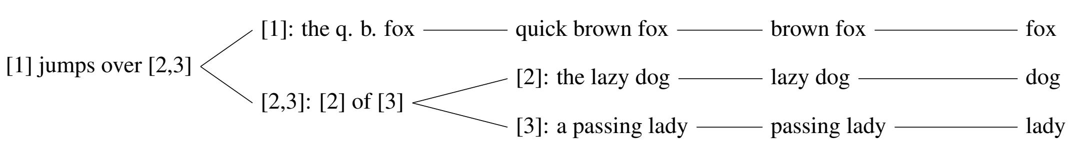
(ii) the **modifier words**, which alter the meaning of individual object words;

(iii) the **interaction words**, which connect the meaning of distinct object words.

We fix a set of **object categories**, both lexical and phrasal, and we declare the **object words** to be those with lexical category in this set (let the set of object words be denoted by $ObjW$). Concretely, we will take the object categories to be the lexical categories N and Pron[4], and the phrasal categories NP and S.

The reason which induced us to include the phrasal category S of sentences into the list of object categories goes as follows. In this work, everything will revolve around a semantic space $\mathscr{H}$ for objects, with no semantic space for sentences. From the point of view of the model presented here, the noun phrase *a jumping fox* and the sentence *a fox jumps* will be treated equally[5], in the sense that both will be taken to be statements about some generic fox which jumps. While the elimination of sentences might seem rather radical, it is consistent with the belief in a world made entirely of interacting objects.

Following this view, we define the **objects** of a sentence $\underline{s}$ (denote their set by $\mathscr{O}[\underline{s}]$) to be the subtrees $T \leq T(\underline{s})$ having an object category as root node; in particular, object words (the leaves of $T(\underline{s})$, seen as singleton subtrees) will be objects of the sentence. We also denote by $\widehat{W}[T]$ the set of word instances in $\underline{s}$ spanned by an object $T \leq T(\underline{s})$, and by $W[T]$ the corresponding set of words $\left\{\overline{w} \in W \times \mathrm{Cat_L} \,\middle|\, \overline{w} = s_j \text{ and } (\underline{s}, j) \in \widehat{W}[T]\right\}$. The set of objects $\mathscr{O}[\underline{s}]$ is a partially ordered set $(\mathscr{O}[\underline{s}], \leq)$ under the subtree ordering $\leq$ inherited from $T(\underline{s})$; in fact, it has the structure of a tree (as long as we assume that the root of $T(\underline{s})$ is always of object category). In the following figure, the poset of objects for sentence (2.1) is shown. Some objects are given labels (numerals in square brackets) for ease of representation.



(2.2)

[2] Additional functional annotations might be required to adapt this model to CCG.

[3] The functions below are *defined* in the remainder of this section. They are not necessarily related to other works.

[4] This is not entirely satisfactory. Treatment of personal pronouns in our framework will certainly require additional efforts.

[5] Assuming that *jumping* and *jumps* would be identified by lemmatisation of the corpus.

Given an object $T \in \mathscr{O}[\underline{s}]$, we define its **completion** $\overline{T}$ to be the largest $T' \geq T$ such that every $T'' \leq T'$ intersects T. In the example (2.2) above, we have $\overline{dog}$ = *the lazy dog*, $\overline{lady}$ = *a passing lady*, and $\overline{fox}$ = *the quick brown fox*. The completion of an object T in a sentence is the largest object in the sentence which contains T and does not contain any object disjoint from T: it completes T to its most detailed description in $\underline{s}$ not involving interactions with independent objects.

We have seen that the objects of sentence are organised in a hierarchy of increasing detail: we wish a modifier word (instance) for an object T to be one which applies to T alone, making it more specific in the sentence. Following this intuition, we define the set $M[T]$ of **modifiers/modifier words** of T to be

$$M[T] := \widehat{W}\left[\overline{T}\right] - \widehat{W}[T] \tag{2.3}$$

In the example (2.2) above, the modifiers of *brown fox* are $\{the, quick\}$, specifying that the object *brown fox* in the sentence is a specific one, and quick as well; similarly, the modifiers of $lady$ are $\{a, passing\}$, specifying that the object $lady$ in the sentence is a generic one, and is passing by. It should be mentioned that this approach is intrinsically intersective in nature: this will reflect in a modelling of modifier words as commuting unary operators[6].

We characterised modifier words for an object T to be those word (instances) which apply to it alone, making it more specific. Dually, we wish to characterise interaction words for two objects T, T' to be those words which are <u>needed</u> to put T and T' in relation in the sentence; in particular, modifiers of T and T' should not appear as interactions. Given two objects $T, T' \in \mathscr{O}[\underline{s}]$, we write $T \vee T'$ to denote the **join** (or most recent common ancestor) of T and T' in the tree $\mathscr{O}[\underline{s}]$. In the example (2.2) above, $dog \vee lady$ = *the lazy dog of a passing lady*, while both $fox \vee lady$ and $fox \vee dog$ are the entire sentence (and so is $fox \vee dog \vee lady$). We shall henceforth require that every object in every sentence of the corpus spans at least an object word: under this assumption, it is always true that a sentence is the completion of the join of all its object words. Following the intuition given above, we define the set $I[T, T']$ of **interactions/interaction words** between T and T' to be

$$I\left[T, T'\right] := \widehat{W}\left[T \vee T'\right] - \widehat{W}\left[\overline{T}\right] - \widehat{W}\left[\overline{T'}\right] \tag{2.4}$$

It should be noted that, in accordance with the intuition given above for completions, we have that $I[T, T'] = I\left[\overline{T}, \overline{T'}\right]$ (because $T \vee T' = \overline{T} \vee \overline{T'}$). In the example (2.2) above, the interactions of *dog* and *lady* form the singleton set $\{of\}$, indicating that the only relationship required to connect *dog* and *lady* is one of possession. Moreover, the interactions of fox and $dog \vee lady$ form the set $\{jumps, over\}$, indicating that *the quick brown fox* and *the lazy dog of a passing lady* are put in relation by an action of jumping and some vertical distance[7]. In order to force compatibility with the left-right ordering of words in pregroup grammars and DisCoCat, we will henceforth set $I[T, T'] := \emptyset$ unless T appears on the left of T' (as disjoint subtrees of $T(\underline{s})$).

We conclude this section by recapping the fundamentals of our toy model for syntax. We have fixed a set of *object categories*, and defined *object words* as words (word instances, to be more precise) from those categories. Based entirely on the respective constituent structure tree, we have given a definition of *objects* in a sentence (subsuming object words are a special case), of *modifier words* (extending objects to more detailed objects) and of *interaction words* (connecting distinct objects).

[6]We observe in passing that non-intersective, contextual behaviour in linguistics requires, at the level of semantics, the same operational features required by contextuality in quantum theory, namely non-commutativity of operators.

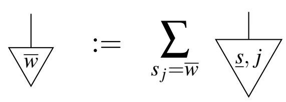
[7]One of the main disadvantages of the intersective approach used here is the issue of phrasal verbs in the English language. A more sophisticated approach, with a more sophisticated grammar, might declare unbreakable phrasal units to cope with this issue (and with some more general instances of non-intersective modifiers).

3 The categorical compositional distributional model

In this section we turn the toy model of semantics defined in the previous section[8] into a categorical compositional distributional model. We fix an involutive[9] commutative semiring $(R,^\star)$, and consider the category R-Mod of finite-dimensional free R-semimodules. The objects of this category are all in the form R^X for some finite set X, and morphisms $R^X \to R^Y$ can be represented as matrices in $R^{Y\times X}$, exactly as in the real and complex cases (with $R = \mathbb{R}$ and $R = \mathbb{C}$ respectively). The category R-Mod is a dagger symmetric monoidal category, with dagger and tensor product defined exactly as in the real and complex cases. Similarly, it is a dagger compact category: as a consequence, it is a suitable model category for categorical compositional distributional semantics. For sake of simplicity, we will restrict ourselves to the case $R = \mathbb{N}$, with the trivial involution $n^\star := n$. A more sophisticated choice of semiring and involution would give us additional freedom in the modelling of words as vectors and R-linear operators: additional semiring elements could be used to signal polarity, modality and/or inflection of word instances; we leave this to future work. It should also be noted that the semantic model presented here is a free one, with a basis ranging over all instances of words in the corpus. Future work will see the development of categorically-sensible compression techniques, both for the purposes of real-world implementation and to enable the emergence of richer semantics from the restriction of available degrees of freedom.

3.1 The distributional part

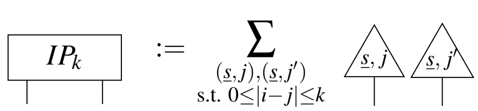
Consider the set $X := \bigcup_{\underline{s}\in\mathscr{C}}\{\underline{s}\} \times \left\{ j \in \{1,...,N_{\underline{s}}\} \,\middle|\, s_j \in ObjW \right\}$ of all instances of object words in sentences of the corpus, and define the **semantic space for objects** $\mathscr{H}$ to be the finite-dimensional free R-semimodule R^X. The R-semimodule $\mathscr{H}$ comes with a **standard orthonormal basis** $\big(|\underline{s}, j\rangle\big)_{(\underline{s},j)\in X}$ indexed by object word instances. To each object word $\overline{w} \in ObjW$ we associate a **semantic vector** $|\overline{w}\rangle$ in $\mathscr{H}$, defined as follows on the standard orthonormal basis:



(3.1)

An object word $\overline{w}$ is then modelled as the indicator function of its instances in the corpus. In particular, different words are assigned orthogonal vectors, and the squared norm $\langle\overline{w}|\overline{w}\rangle \in R$ of a word counts the total number of occurrences of $\overline{w}$ throughout the corpus. In other approaches, the inner product of the vectors associated with two words is used to encode some notion of semantic distance between them. One such notion, that of co-occurrence (within a window of $2k+1$ object words, with k fixed), can be easily recovered in the framework presented here (and so can many other linear notions of distance).

Recall that the inner product can be obtained as $\langle\overline{w}|\overline{w}\rangle = \cap_{\mathscr{H}} \cdot \big(|\overline{w}\rangle^\star \otimes |\overline{w}\rangle\big)$, where the bilinear map $\cap : \mathscr{H}\otimes\mathscr{H} \to R$ is the usual cap[10] given by the dagger compact structure. The familiar co-occurrence picture can be recovered by considering a different bilinear map $IP_k : \mathscr{H}\otimes\mathscr{H} \to R$, which counts the co-occurrences $IP_k \cdot \big(|\overline{w}\rangle^\star \otimes |\overline{w}'\rangle\big) \in R$ of two object words $\overline{w}, \overline{w}'$ in windows of $2k+1$ object words within the same sentence:



(3.2)

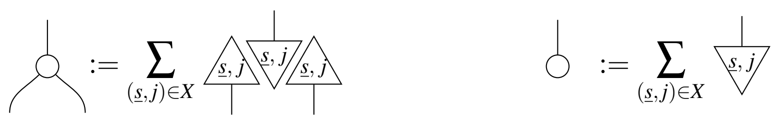
[8] Consisting of object words, modifier words, and interaction words, together with their hierarchical structure.

[9] I.e. a commutative semiring with a chosen involution *, with similar axioms to conjugation in the semiring/field $\mathbb{C}$.

[10] The dagger compact category R-Mod has self-dual objects, where the conjugate $|\psi\rangle^\star \in R^Y$ of a vector $|\psi\rangle \in R^Y$ is obtained by application of the involution * to each coordinate of $|\psi\rangle$ in the standard orthonormal basis of R^Y.

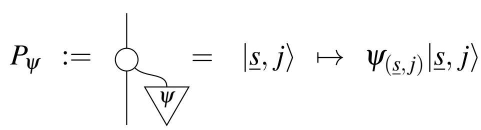
3.2 The compositional part

Now that object words have been associated to vectors in the semantic space $\mathscr{H}$, we will model modifiers and interactions as operators. When $\mathscr{H}$ is very high-dimensional (as is certainly the case here), operators $\mathscr{H} \to \mathscr{H}$ can be extremely expensive to concretely work with: as a consequence, one often restricts the attention to a more convenient subclass of operators, obtained by using Frobenius algebras. No matter what semiring R we choose, the standard orthonormal basis of $\mathscr{H}$ always gives rise to a special commutative †-Frobenius algebra ($\cdot$, $\cdot$, $\cdot$, $\cdot$) in the following way:



(3.3)

One then considers the family of operators $\mathscr{H} \to \mathscr{H}$ taking the following form, which can be efficiently[11] represented by means of vectors in $\mathscr{H}$:

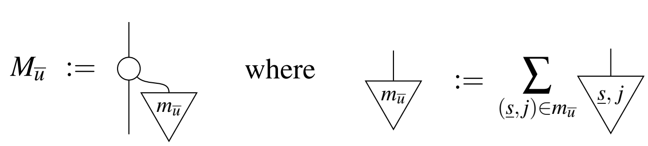


(3.4)

Operators in this form can be easily manipulated using the algebra operations: for example, if $|\psi \odot \phi\rangle :=$ $\cdot$ $\cdot (|\psi\rangle \otimes |\phi\rangle)$ denotes the algebra multiplication, then composition of operators in the form above is given by $P_\phi \circ P_\psi = P_{\psi \odot \phi}$ (using associativity of the algebra multiplication).

A particularly interesting class of operators $\mathscr{H} \to \mathscr{H}$ is given by the **projectors**, the self-adjoint idempotent operators. It is easy to show[12] that an operator in the form (3.4) above is a projector if and only if all the coordinates $\psi_{(\underline{s},j)}$ are self-conjugate idempotents in the semiring R. In the case of fields (such as $R = \mathbb{R}$ or $R = \mathbb{C}$), or semirings with cancellation (such as $R = \mathbb{Z}$ or $R = \mathbb{N}$), the only idempotents in R are 0 and 1; however, more general semirings will have more (self-conjugate) idempotents, giving us more projectors.

Modifiers were defined to be words which alter the meaning of individual object words by making it more specific: following this intuition, we model them as projectors $\mathscr{H} \to \mathscr{H}$. Given a non-object word $\overline{u}$, we define the associated **modifier** $M_{\overline{u}} : \mathscr{H} \to \mathscr{H}$ to be the projector over the subspace given by all the objects which $\overline{u}$ modifies in the corpus:



(3.5)

and we defined $m_{\overline{u}} := \{(\underline{s}, j) \in X \,|\, \exists T \in \mathscr{O}[\underline{s}] \text{ with } \overline{u} \in M[T] \text{ and } (\underline{s}, j) \leq T\}$ to be the set of instances of object words which appear in objects modified by $\overline{u}$. A modifier $M_{\overline{u}}$ sends the semantic vector $|\overline{w}\rangle$ of an object word $\overline{w}$ to the indicator function $M_{\overline{u}}|\overline{w}\rangle$ of all the instances of $\overline{w}$ appearing in objects modified by $\overline{u}$.

[11] If $\mathscr{H}$ is M-dimensional, then these operators admit an M-dimensional representation, compared with the M^2-dimensional representation required in general by operators $\mathscr{H} \to \mathscr{H}$.

[12] Observing that $(\psi \odot \phi)_{(\underline{s},j)} = \psi_{(\underline{s},j)} \phi_{(\underline{s},j)}$ in R.

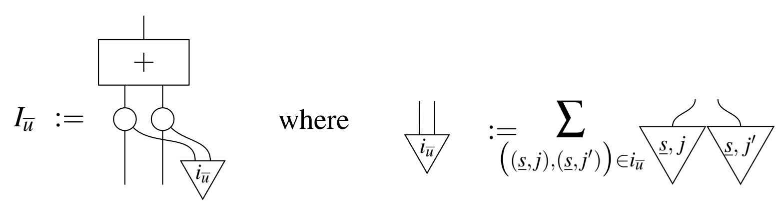
In their formulation as projectors, modifiers come with some interesting logical structure (in the language of quantum theory, the structure of a commutative algebra of projectors). If the semiring R satisfies the additive cancellation law (i.e. if $a = b$ whenever $a + c = b + c$ for some $c \in R$), then it is possible to define a partial subtraction operation on elements by setting $b - a := c$ whenever a c exists such that $b = a + c$ (if it exists, it is necessarily unique). In $\mathbb{N}$, this returns the usual subtraction (and similarly in $\mathbb{Z}$ with full-fledged additive inverses). When R satisfies the additive cancellation law, modifiers come with natural logical operations:

(i) if $\overline{u}$ and $\overline{v}$ are two non-object words, then $M_{\overline{u}} \circ M_{\overline{v}}$ (which always equals $M_{\overline{v}} \circ M_{\overline{u}}$ since (⅄ , ⫯ , Y , ♀) is commutative) can equivalently be obtained as $P_{m_{\overline{u}} \odot m_{\overline{v}}}$, and is a well-defined projector[13]. It projects onto the subspace given by all object word instances appearing in objects modified by <u>both</u> $\overline{u}$ and $\overline{v}$. We will also write this as $M_{\overline{u}} \odot M_{\overline{v}}$.

(ii) if $\overline{u}$ and $\overline{v}$ are two non-object words, then $M_{\overline{u}} + M_{\overline{v}} - M_{\overline{u}} \odot M_{\overline{v}}$ is always a well defined projector. It projects on the subspace given by all object word instances appearing in objects modified by <u>at least one</u> of $\overline{u}$ and $\overline{v}$. We will write this as $M_{\overline{u}} \oplus M_{\overline{v}}$.

(iii) if $\overline{u}$ is a non-object word, then $id_{\mathscr{H}} - M_{\overline{u}}$ is always a well defined projector. It projects onto the subspace given by all object word instances appearing in objects <u>not</u> modified by $\overline{u}$. We will write this as $1 \ominus M_{\overline{u}}$.

Interactions were defined to be words which connect the meaning of different object words within the same sentence: we model them as operators $\mathscr{H} \otimes \mathscr{H} \to \mathscr{H}$. Given a word $\overline{u}$ not of object category, we define the associated **interaction** $I_{\overline{u}} : \mathscr{H} \otimes \mathscr{H} \to \mathscr{H}$ to be the following binary operation:



(3.6)

and we defined $i_{\overline{u}} := \left\{ \left((\underline{s}, j), (\underline{s}, j')\right) \in X \times X \,\middle|\, \exists T, T' \in \mathscr{O}[\underline{s}] \text{ with } \overline{u} \in I[T, T'] \text{ and } (\underline{s}, j) \leq T, (\underline{s}, j') \leq T' \right\}$ to be the set of pairs of instances which appear in objects put into relation by $\overline{u}$. The interaction $I_{\overline{u}}$ is obtained as a projector on $\mathscr{H} \otimes \mathscr{H}$, selecting pairs of instances in $i_{\overline{u}}$, followed by the linear operator $\boxplus : \mathscr{H} \otimes \mathscr{H} \to \mathscr{H}$ defined as follows:

$$\boxplus := \left(id_{\mathscr{H}} \otimes \text{♀} + \text{♀} \otimes id_{\mathscr{H}} \right) = |\underline{s}, j\rangle \otimes |\underline{s}', j'\rangle \mapsto |\underline{s}, j\rangle + |\underline{s}', j'\rangle \tag{3.7}$$

The interaction $I_{\overline{u}}$ sends a pair $\overline{w}, \overline{w}'$ of object words to the vector $|\psi\rangle$ spanning all object words appearing in objects containing instances of $\overline{w}$ and $\overline{w}'$ and related by $\overline{u}$:

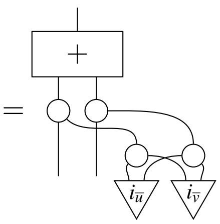
$$I_{\overline{u}} \cdot \left(|\overline{w}\rangle \otimes |\overline{w}'\rangle \right) \;=\; \sum_{\left((\underline{s},j),(\underline{s},j')\right) \in i_{\overline{u}}(\overline{w},\overline{w}')} \left(|\underline{s}, j\rangle + |\underline{s}, j'\rangle \right) \tag{3.8}$$

where we defined $i_{\overline{u}}(\overline{w}, \overline{w}') := \left\{ \left((\underline{s}, j), (\underline{s}, j')\right) \in i_{\overline{u}} \,\middle|\, \overline{w} = s_j \text{ and } \overline{w}' = s_j \right\}$. Note that the instances $(\underline{s}, j)$ and $(\underline{s}, j')$ appearing in the sum of Equation (3.8) are necessarily distinct even when $\overline{w} = \overline{w}'$.

[13] Frobenius algebras thus play a very active role in information flow within the sentence, combining modifier words together into larger modifiers. Later on, they will also be seen to combine interaction words together into larger interactions.

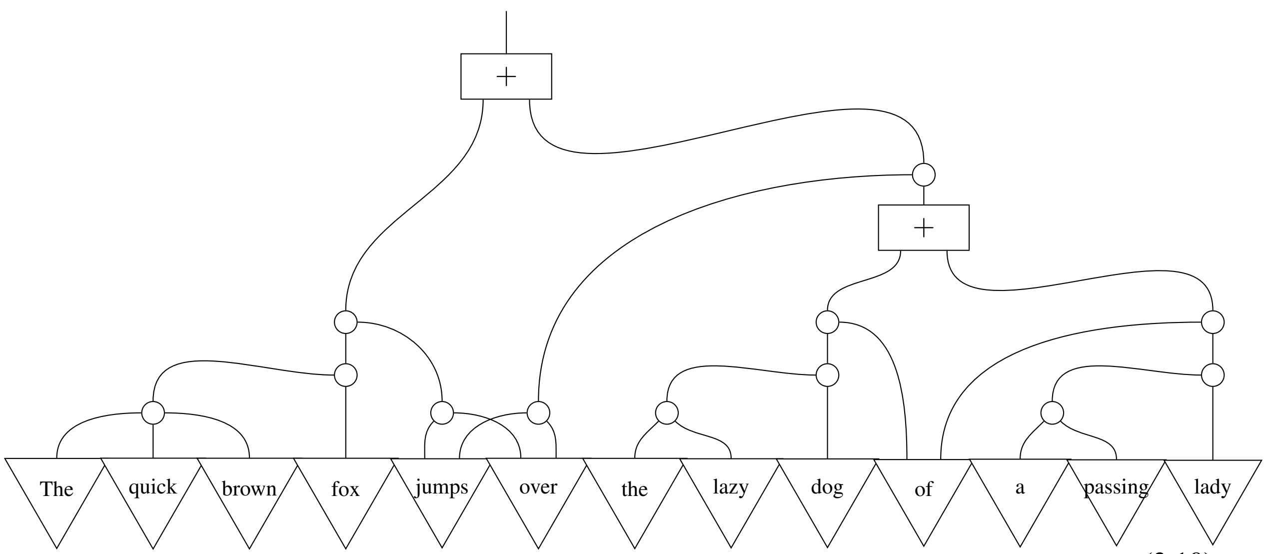
Contrary to modifiers, interactions associated with two non-object words $\overline{u}, \overline{v}$ cannot be composed directly to obtain an interaction, because of the $\boxplus$ map compressing $\mathscr{H} \otimes \mathscr{H}$ to $\mathscr{H}$. However, the projector parts of the interactions do compose, and the joint interaction $I_{\overline{u}} \odot I_{\overline{v}}$ can be defined by using the Frobenius multiplication $\curlywedge \otimes \curlywedge$ on $\mathscr{H} \otimes \mathscr{H}$:

$I_{\overline{u}} \odot I_{\overline{v}} :=$



(3.9)

Finally, the following figure presents the full semantic vector associated with the original sentence from example (2.1). Associativity and commutativity of the Frobenius algebra multiplication has been used to group modifiers and interactions together, improving readability.



(3.10)

4 Future work

We look forward to improve and extend the model in the following directions. First of all, a more sophisticated choice of commutative semiring R would give us additional semantic degrees of freedom: more idempotents could be used to identify the different roles played by words in interactions (e.g. by introducing a *subject* idempotent and an *object* idempotent), while a self-inverse element could be used to distinguish positive occurrences of object words from negated ones. Second, the category R-Mod admits a CPM construction, which can be used to model ambiguous semantics. Finally, the construction given here is a *free* one, with basis ranging over all instances of words in the corpus: for efficient real-world applications, a suitable compressed construction should be devised. Also, the case of conjunctions needs to be fully treated, and a better modelling of personal pronouns is deemed necessary.

Acknowledgements The author would like to thank Bob Coecke for suggestions, comments and useful discussions, as well as Sukrita Chatterji and Nicolò Chiappori for their continued support. Funding from EPSRC and Trinity College is gratefully acknowledged.

References

[1] Samson Abramsky & Bob Coecke (2009): *Categorical quantum mechanics*. Handbook of quantum logic and quantum structures: quantum logic, doi:10.1016/B978-0-444-52869-8.50010-4.

[2] Esma Balkir, Mehrnoosh Sadrzadeh & Bob Coecke (2016): *Distributional sentence entailment using density matrices*. In: Topics in Theoretical Computer Science, Springer, doi:10.1007/978-3-319-28678-5_1.

[3] Noam Chomsky (1995): *The minimalist program*. Cambridge University Press.

[4] Bob Coecke, Mehrnoosh Sadrzadeh & Stephen Clark (2010): *Mathematical foundations for a compositional distributional model of meaning*.

[5] John Rupert Firth (1957): *Papers in linguistics, 1934-1951*. Oxford University Press.

[6] Zellig Harris (1968): *Mathematical structures of language*. Interscience Publishers John Wiley & Sons.

[7] Dimitri Kartsaklis & Mehrnoosh Sadrzadeh (2015): *A Frobenius model of information structure in categorical compositional distributional semantics*. In: Proceedings of the 14th Meeting on the Mathematics of Language, doi:10.3115/v1/W15-2306.

[8] Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, Stephen Pulman & Bob Coecke (2013): *Reasoning about meaning in natural language with compact closed categories and frobenius algebras*.

[9] Sandra Kübler, Ryan McDonald & Joakim Nivre (2009): *Dependency parsing*. Synthesis Lectures on Human Language Technologies, doi:10.2200/S00169ED1V01Y200901HLT002.

[10] Joachim Lambek (1958): *The mathematics of sentence structure*. The American Mathematical Monthly, doi:10.2307/2310058.

[11] Joachim Lambek (1999): *Type grammar revisited*. Logical Aspects of Computational Linguistics, doi:10.1007/3-540-48975-4_1.

[12] Robin Piedeleu, Dimitri Kartsaklis, Bob Coecke & Mehrnoosh Sadrzadeh (2015): *Open system categorical quantum semantics in natural language processing*. In: Proceedings of the 6th Conference on Algebra and Coalgebra in Computer Science, doi:10.4230/LIPIcs.CALCO.2015.270.

[13] Mehrnoosh Sadrzadeh, Stephen Clark & Bob Coecke (2013): *The Frobenius anatomy of word meanings I: subject and object relative pronouns*. Journal of Logic and Computation, doi:10.1093/logcom/ext044.

[14] 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.

[15] Peter Selinger (2007): *Dagger compact closed categories and completely positive maps*. Electronic Notes in Theoretical Computer Science, doi:10.1016/j.entcs.2006.12.018.

[16] Mark Steedman (2000): *The syntactic process*. MIT Press.